PATENT ABSTRACTS OF JAPAN

(11)Publication number : **2001-188555**

(43)Date of publication of application : **10.07.2001**

(51)Int.Cl. **G10L 15/00**
B25J 13/00
G06T 1/00
G06T 7/20
G10L 15/24

(21)Application number : **11-375773**

(22)Date of filing : **28.12.1999**

(71)Applicant : **SONY CORP**

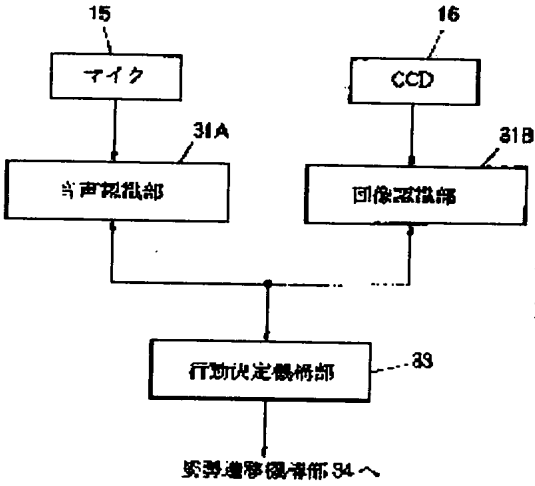
(72)Inventor : **YAMASHITA JUNICHI**
OGAWA HIROAKI
HONDA HITOSHI
HELMUT LUCKE
TAMARU EIJI
FUJITA YAEKO

(54) DEVICE AND METHOD FOR INFORMATION PROCESSING AND RECORDING MEDIUM

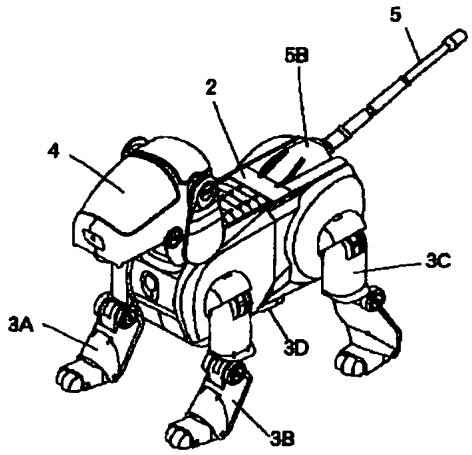
(57)Abstract:

PROBLEM TO BE SOLVED: To provide a robot-performing operation which is rich in variety.

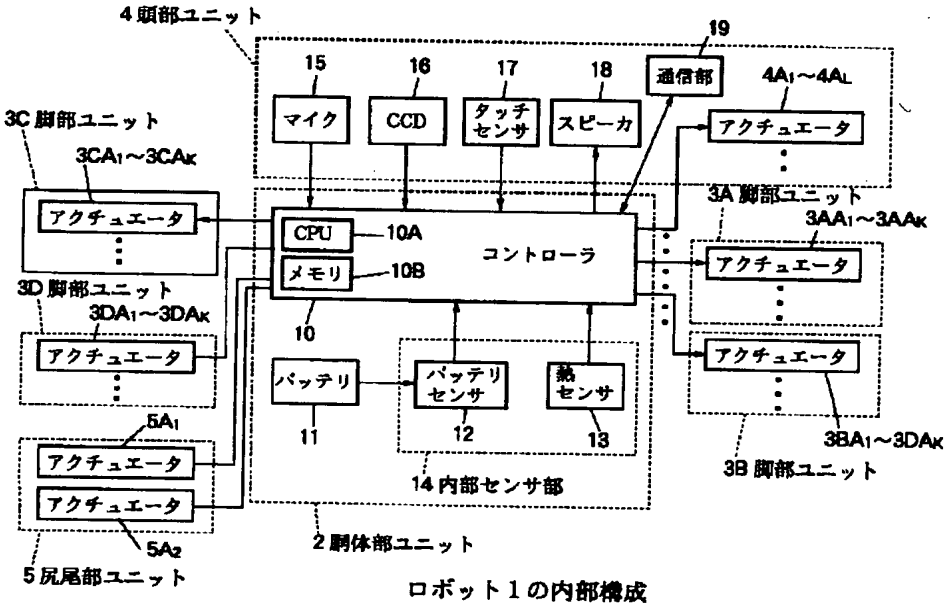
SOLUTION: A user's voice which is picked up by a microphone 15 is recognized by a voice recognition part 31A. A gesture of the user photographed by a CCD 16 is recognized by an image recognition part 31B. An action- determining mechanism part 33 determines the operation of the robot by using the voice information outputted from the voice recognition part 31A and the image information outputted from the image recognition part 31B.



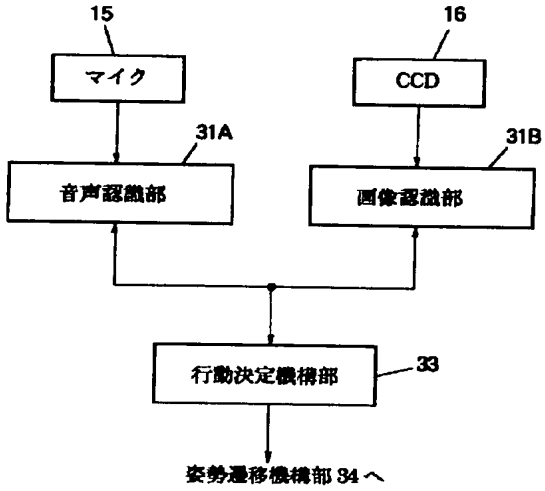
DRAWINGS



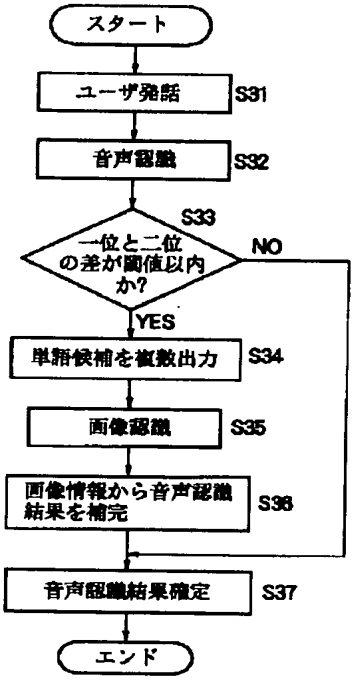
[Drawing 1] ペットロボット１の外観構成



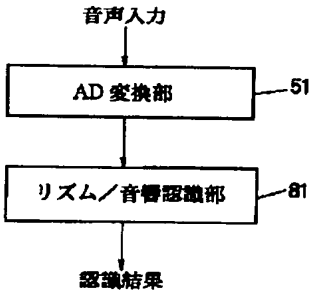
[Drawing 2] ロボット１の内部構成



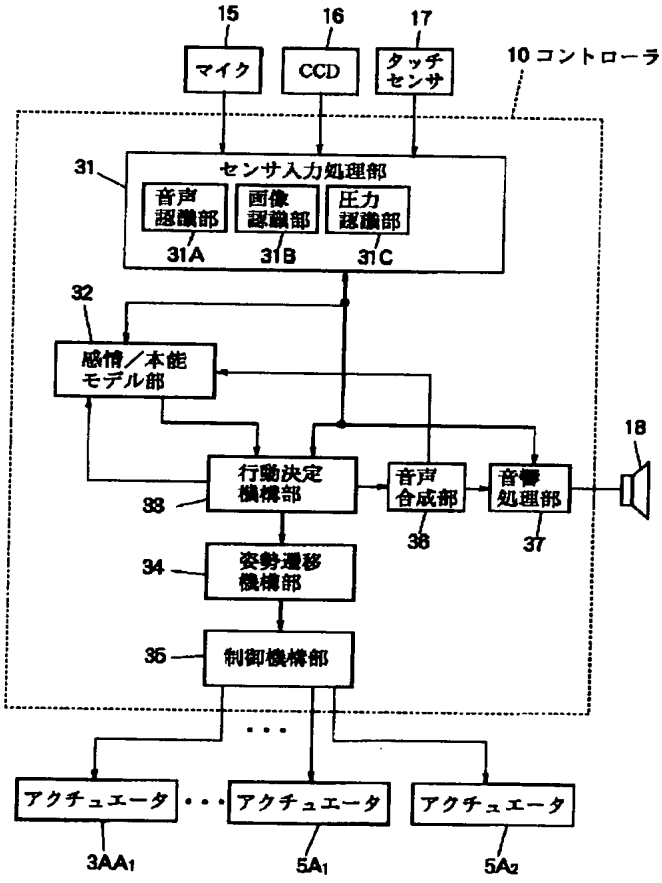
[Drawing 4]



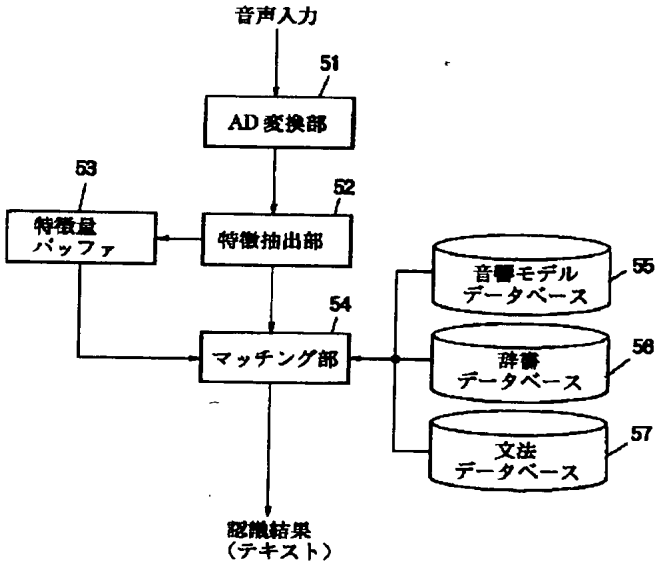
[Drawing 13]



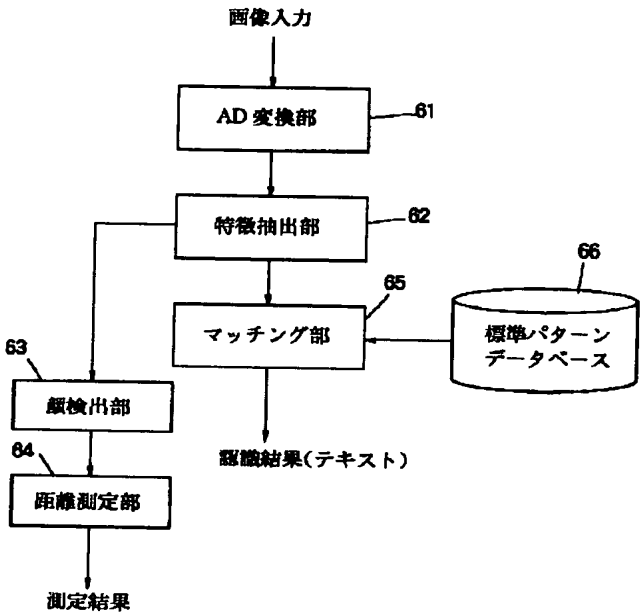
[Drawing 17] 音声認識処理部 A



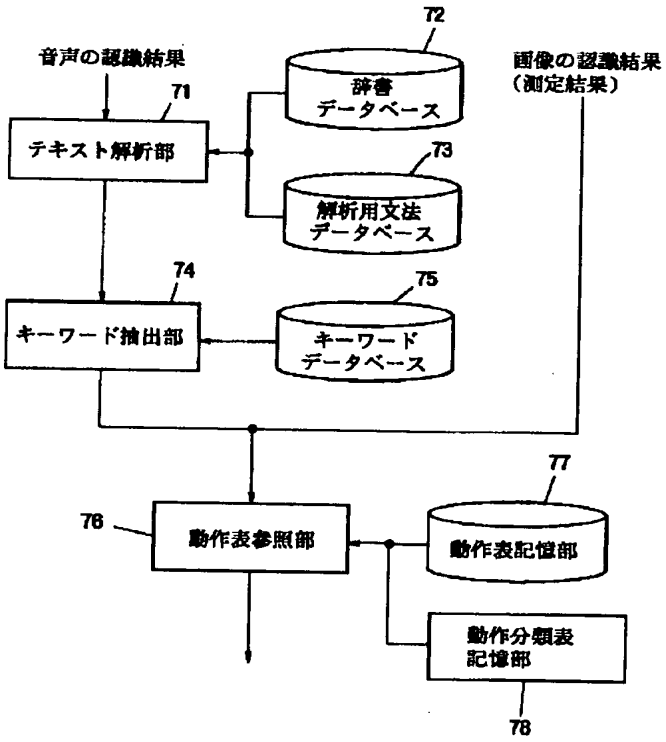
[Drawing 3]



[Drawing 5] **音声認識部 31A**



[Drawing 6] **画像認識部 31B**



[Drawing 7]

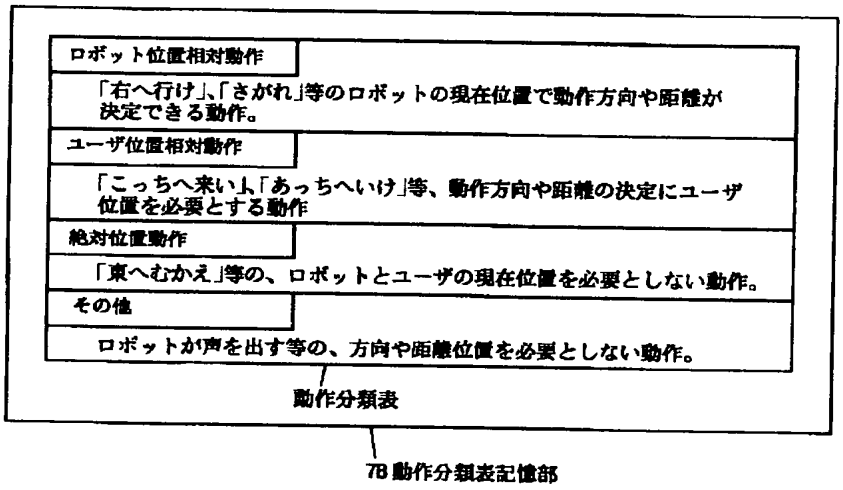
行動決定機構部 33

画像の認識結果	付帯情報	音声の認識結果		動作
手招き	ユーザの位置	こっち来い等 ロボットの名前(名前を呼ぶ)		ユーザに近づく、離れる、無視
ゆびを指す	指の方向	感嘆表現(あっ、おやっ等)		指の差している方向を向く
		～を取って	指示代名詞	指の指す方向にある対象物を切り出し指示代名詞を補完
			他の名前	ゆびの指す方向で物を検索する
握手		挨拶		手を前に出す
手を振る		別れの挨拶(バイバイ等)		手を振る、ユーザーから離れる、電源を切る
		挨拶(おーい等呼びかけ)		近づく、挨拶
		なし		手を振る
なし		呼びかけ		ユーザーを探す

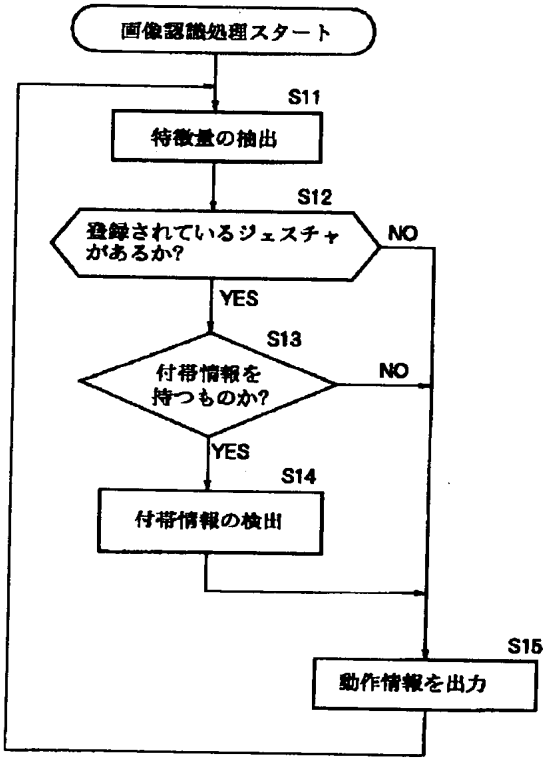
動作表

[Drawing 8]

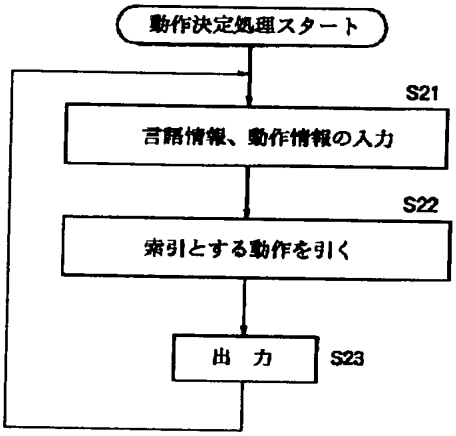
77 動作表記憶部



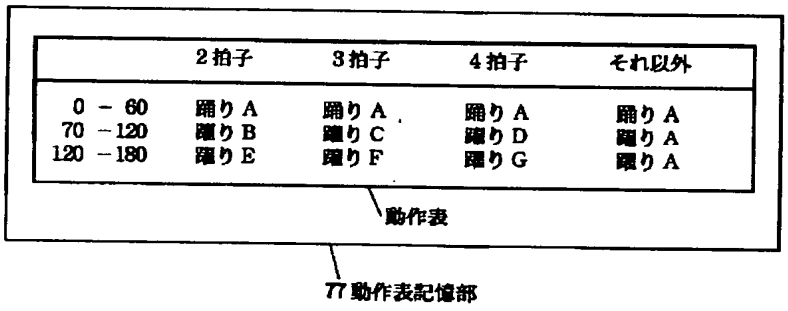
[Drawing 9]



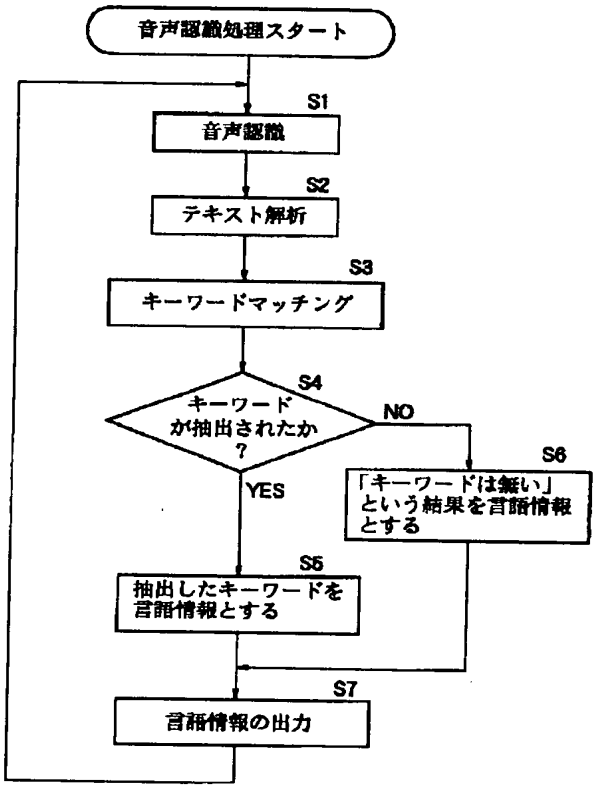
[Drawing 11]



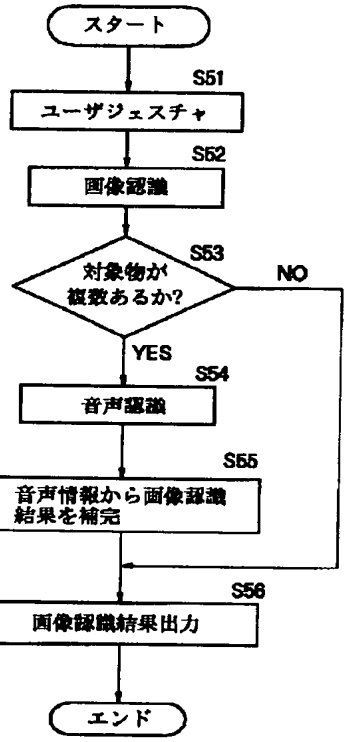
[Drawing 12]



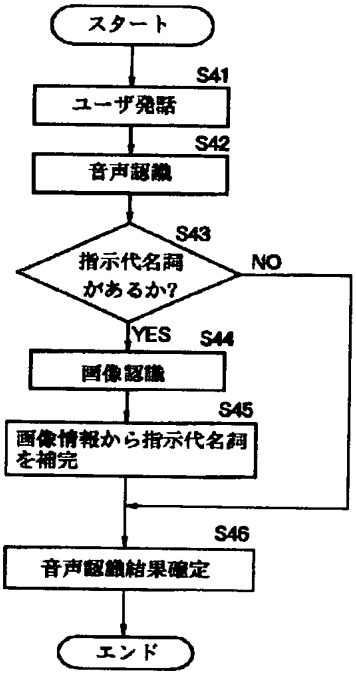
[Drawing 18]



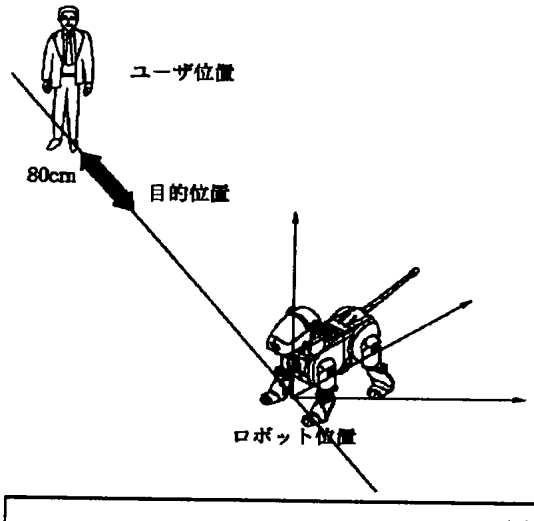
[Drawing 10]



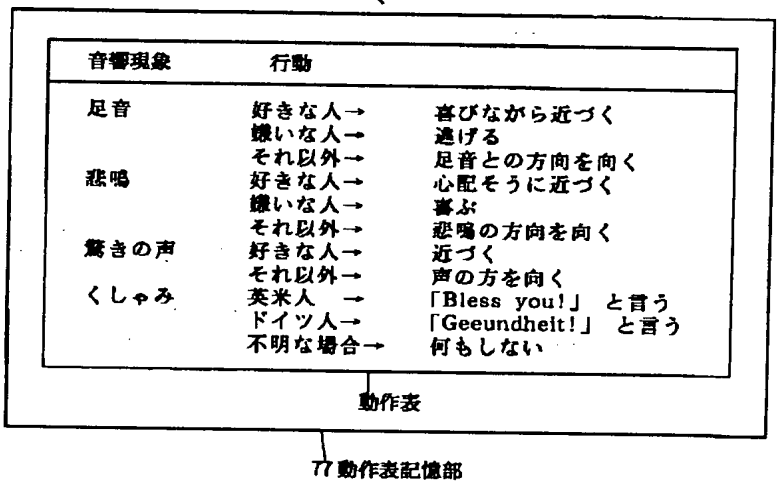
[Drawing 15]



[Drawing 14]



[Drawing 16]

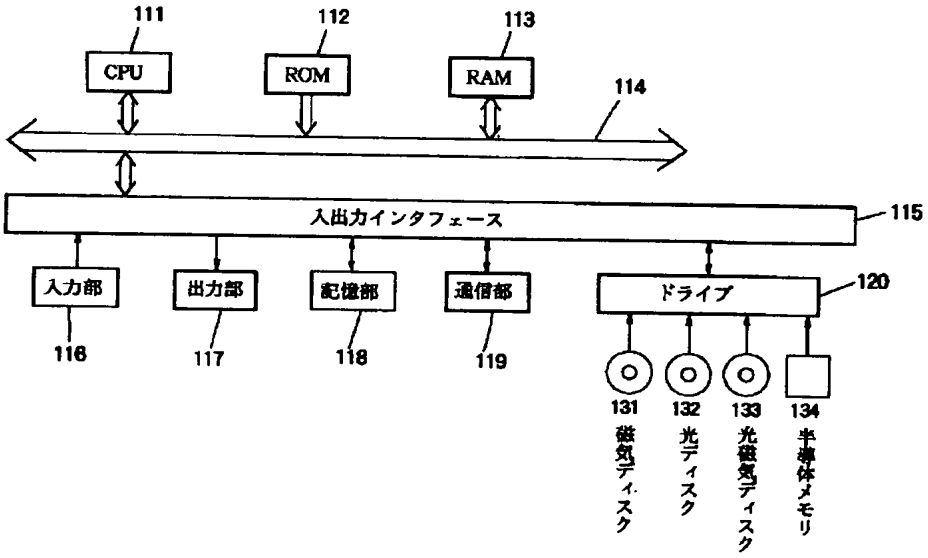


音響現象	行動	
足音	好きな人→	喜びながら近づく
	嫌いな人→	逃げる
	それ以外→	足音との方向を向く
悲鳴	好きな人→	心配そうに近づく
	嫌いな人→	喜ぶ
	それ以外→	悲鳴の方向を向く
驚きの声	好きな人→	近づく
	それ以外→	声の方を向く
くしゃみ	英米人 →	「Bless you!」 と言う
	ドイツ人→	「Geeundheit!」 と言う
	不明な場合→	何もしない

動作表

77 動作表記憶部

[Drawing 19]



[Drawing 20]

DETAILED DESCRIPTION

[Detailed Description of the Invention]

[0001]

[Field of the Invention] This invention is used for a robot which determines actuation as an information processor and an approach, and a list especially about a record medium using speech information and image information, and relates to a record medium at a suitable information processor and a suitable approach, and a list.

[0002]

[Description of the Prior Art] Conventionally, as a toy etc., if press actuation of the touch switch is carried out, many robots (a sewing basis-like thing is included) which output composite tone, robots which can enjoy a false conversation of recognizing a user's utterance and returning a response sentence are produced commercially.

[0003] Moreover, an image is picturized, image recognition is performed based on the picturized image, circumstantial judgment is performed, and the robot which acts autonomously is produced commercially.

[0004]

[Problem(s) to be Solved by the Invention] However, the recognition result by speech recognition had the technical problem that incorrect recognition will be carried out, when a user's utterance was not performed clearly. Moreover, the technical problem that the object which a demonstrative pronoun directs may be unable to be recognized occurred the case of the utterance containing an ambiguous word, for example, the utterance containing a demonstrative pronoun.

[0005] Moreover, it was difficult for the robot which mentioned above to perform autonomous actuation only depending on one information on voice or an image, and to perform autonomous actuation using the information on voice and an image both.

[0006] By making this invention in view of such a situation, and using the information on both voice and an image, more positive speech recognition is performed and it aims at providing a user with the actuation which was rich in variety.

[0007]

[Means for Solving the Problem] An information processor according to claim 1 is characterized by

including a decision means to opt for actuation of a robot at least using one side among the recognition result by speech recognition means to recognize voice, image recognition means to recognize an image, and the speech recognition means, or the recognition result by the image recognition means.

[0008] A maintenance means to hold the table on which the relation of a robot of operation determined as a meaning by the recognition result by said speech recognition means, the recognition results by the image recognition means, and those recognition results was described can be included further.

[0009] When the recognition result by the voice means is not determined as a meaning, using the recognition result by the image recognition means, said decision means is determined as a meaning and can opt for actuation of a robot using the determined recognition result.

[0010] Said decision means can opt for actuation of a robot using the recognition result determined by the meaning using the recognition result by the speech recognition means, when two or more objects exist in the image which an image recognition means recognizes.

[0011] Said image recognition means detects the direction to which the part beforehand set up among a user's finger, the face, the eye, and the jaw points, and can recognize the image located in the direction.

[0012] Including further a storage means to memorize the data about the gesture which a user performs, by recognizing a user's image, an image recognition means detects the gesture memorized by the storage means, and can make a recognition result the detected result.

[0013] It is characterized by also using the measurement result by the measurement means for a decision means including a measurement means to measure the distance between a user and self, and opting for actuation of a robot further, from the magnitude of a user's face detected by detection means to detect a user's face, and the detection means.

[0014] Said speech recognition means detects the rhythm contained in an environmental sound, and can make it a recognition result.

[0015] From an environmental sound, said speech recognition means detects a sound phenomenon, and can make it a recognition result.

[0016] The information processing approach according to claim 10 is characterized by including the decision step which opts for actuation of a robot at least using one side among the recognition result by processing of the speech recognition step which recognizes voice, the image recognition step which recognizes an image, and a speech recognition step, or the recognition result by processing of an image recognition step.

[0017] The program of a record medium according to claim 11 is characterized by including the decision step which opts for actuation of a robot at least using one side among the recognition result by processing of the speech recognition step which recognizes voice, the image recognition step which recognizes an image, and a speech recognition step, or the recognition result by processing of an image recognition step.

[0018] In an information processor according to claim 1, the information processing approach according to claim 10, and a record medium according to claim 11, voice is recognized, an image is recognized and it opts for actuation of a robot among an audio recognition result or the recognition result of an image at least using one side.

[0019]

[Embodiment of the Invention] Drawing 1 shows the example of an appearance configuration of the gestalt of 1 operation of the robot which applied this invention, and drawing 2 shows the example of an electric configuration.

[0020] The robot consists of gestalten of this operation by connecting the head unit 4 and the tail section unit 5 with the front end section and the back end section of the idiosoma unit 2, respectively while considering as the thing of a dog configuration and connecting the leg units 3A, 3B, and 3C and 3D with front and rear, right and left of the idiosoma unit 2, respectively.

[0021] The tail section unit 5 is pulled out free [a curve or rocking] with two degrees of freedom from base section 5B prepared in the top face of the idiosoma unit 2. The controller 10 which controls the whole robot, the dc-battery 11 used as a robot's source of power, the internal sensor section 14 which becomes a list from

the dc-battery sensor 12 and the heat sensor 13 are contained by the idiosoma unit 2.

[0022] The microphone (microphone) 15 which is equivalent to a "lug" at the head unit 4, the CCD (Charge Coupled Device) camera 16 equivalent to a "eye", the touch sensor 17 equivalent to a tactile sense, the loudspeaker 18 equivalent to "opening", etc. are arranged in the predetermined location, respectively.

[0023] Leg unit 3A thru/or the joint part of each 3D, and leg unit 3A thru/or each 3D and the joining segment of the idiosoma unit 2, In the joining segment of the head unit 4 and the idiosoma unit 2, and a list, to the joining segment of the tail section unit 5 and the idiosoma unit 2 As shown in drawing 2 , actuator 3AA1 thru/or 3AAK(s), 3BA1 or 3BAK(s), 3CA1 or 3CAK(s), 3D A1 or 3D AK, four A1 or 4AL(s), five A1, and five A2 are arranged, respectively. By this Each joining segment can have a predetermined degree of freedom, and can rotate it now.

[0024] The microphone 15 in the head unit 4 collects the voice (sound) of a perimeter including the utterance from a user, and sends out the acquired sound signal to a controller 10. CCD camera 16 picturizes a surrounding situation and sends out the acquired picture signal to a controller 10.

[0025] The touch sensor 17 is formed in the upper part of the head unit 4, detects the pressure received by "it strokes" and the physical influence of "striking" from a user, and sends it out to a controller 10 by making the detection result into a pressure detecting signal.

[0026] The dc-battery sensor 12 in the idiosoma unit 2 detects the residue of a dc-battery 11, and sends out the detection result to a controller 10 as a dc-battery residue detecting signal. The heat sensor 13 detects the heat inside a robot, and sends out the detection result to a controller 10 as a heat detecting signal.

[0027] The controller 10 contains CPU(Central Processing Unit)10A, memory 10B, etc., and performs various kinds of processings by performing the control program memorized by memory 10B in CPU10A. That is, a controller 10 judges existence, such as a surrounding situation, and a command from a user, influence from a user, based on the sound signal given from a microphone 15, CCD camera 16 and a touch sensor 17, the dc-battery sensor 12, and the heat sensor 13, a picture signal, a pressure detecting signal, a dc-battery residue detecting signal, and a heat detecting signal.

[0028] Furthermore, a controller 10 opts for the continuing action based on this decision result etc. The required thing of actuator 3AA1 thru/or 3AAK(s), 3BA1 or 3BAK(s), 3CA1 or 3CAK(s), 3D A1 or 3D AK, four A1 or 4AL(s), five A1, and five A2 is made to drive based on the decision result. By this The head unit 4 can be made to be able to shake vertically and horizontally, the tail section unit 5 can be moved, or each leg unit 3A thru/or 3D are driven, and it makes it act to walk him around a robot etc.

[0029] Moreover, if needed, a controller 10 generates composite tone, and it is made to supply and output to a loudspeaker 18, or it turns on, switches off or blinks LED (Light Emitting Diode) which was prepared in the location of the "eyes" of a robot and which is not illustrated.

[0030] A robot can take action now autonomously based on a surrounding situation etc. as mentioned above.

[0031] Next, drawing 3 shows the example of a functional configuration of the controller 10 of drawing 2 . In addition, the functional configuration shown in drawing 3 is realized because CPU10A performs the control program memorized by memory 10B.

[0032] A controller 10 accumulates the recognition result of the sensor input-process section 31 which recognizes a specific external condition, and the sensor input-process section 31 etc. It is based on the recognition result expressing the condition of feeling and instinct of feeling / instinct model section 32, and the sensor input-process section 31 etc. It is based on the decision result of the action decision mechanism section 33 which opts for the continuing action, and the action decision mechanism section 33. It consists of the controlling mechanism section 35 which carries out drive control of the posture transition device section 34 which makes a robot actually take action, each actuator 3AA1 or five A1, and five A2, the speech synthesis section 36 which generates composite tone, and the acoustical-treatment section 37 which controls the output of the speech synthesis section 36 in a list.

[0033] Based on a microphone 15, CCD camera 16, the sound signal given from touch sensor 17 grade, a picture signal, a pressure detecting signal, etc., the sensor input-process section 31 recognizes a specific

external condition, the specific influence by the user, the directions from a user, etc., and notifies the condition recognition information that the recognition result is expressed to feeling / instinct model section 32, and the action decision mechanism section 33.

[0034] Namely, it has speech recognition section 31A, and, as for the sensor input-process section 31, speech recognition section 31A performs speech recognition using the sound signal given from a microphone 15 according to the control from the action decision mechanism section 33. And speech recognition section 31A notifies the command and others as the speech recognition result, such as "walk", "lie down", and "pursue a ball", to feeling / instinct model section 32, and the action decision mechanism section 33 as condition recognition information.

[0035] Moreover, it has image recognition section 31B, and, as for the sensor input-process section 31, image recognition section 31B performs image recognition processing using the picture signal given from CCD camera 16. and image recognition section 31B -- the result of the processing -- for example, -- "-- when flat-surface" more than the ****** height in a perpendicular etc. is detected to round red thing" and "ground, an image recognition result, such as "there is a ball" or there "there being a wall", is notified to feeling / instinct model section 32, and the action decision mechanism section 33 as condition recognition information. Moreover, recognition of gesture which a user performs is also performed and the recognition result is notified to the action decision mechanism section 33.

[0036] Furthermore, the sensor input-process section 31 has pressure processing section 31C, and pressure processing section 31C processes the pressure detecting signal given from a touch sensor 17. and when it is beyond a predetermined threshold and a short-time pressure is detected as a result of the processing, pressure processing section 31C It is recognized as "It was struck" (cut by carrying out), it is under a predetermined threshold, and when the pressure of long duration is detected, it is recognized as "It was stroked" (praised) and the recognition result is notified to feeling / instinct model section 32, and the action decision mechanism section 33 as condition recognition information.

[0037] Feeling / instinct model section 32 has managed the feeling model and instinct model expressing a robot's feeling and the condition of instinct, respectively. Based on the condition recognition information from the sensor input-process section 31, the feeling / instinct status information from feeling / instinct model section 32, time amount progress, etc., the action decision mechanism section 33 opts for the next action, and sends it out to the posture transition device section 34 by making into action command information the contents of the action for which it opted.

[0038] The posture transition device section 34 generates the posture transition information for making a robot's posture change into the following posture from a current posture based on the action command information supplied from the action decision mechanism section 33, and outputs this to the controlling mechanism section 35. The controlling mechanism section 35 generates the control signal for driving actuator 3AA1 thru/or five A1, and five A2 according to the posture transition information from the posture transition device section 34, and sends this out to actuator 3AA1 thru/or five A1, and five A2. Thereby, actuator 3AA1 thru/or five A1, and five A2 are driven according to a control signal, and a robot takes action autonomously.

[0039] A robot 1 recognizes a user's voice and gesture and opts for action. From the example of a functional configuration shown in drawing 3 , a user's voice and gesture are recognized and what took out the part for opting for action is shown in drawing 4 . That is, in order to recognize a user's voice and to recognize the gesture of a microphone 15, speech recognition section 31A, and a user, it has CCD16 and image recognition section 31B, and the action decision mechanism section 33 opts for a robot's 1 action by the recognition result obtained from speech recognition section 31A and image recognition section 31B.

[0040] Drawing 5 is drawing showing the detailed configuration of speech recognition section 31A. A user's utterance is inputted into a microphone 15 and the utterance is changed into the sound signal as an electrical signal on a microphone 15. This sound signal is supplied to the AD (Analog Digital) transducer 51 of speech recognition section 31A. In the AD translation section 51, the sound signal which is an analog signal from a

microphone 15 is sampled and quantized, and it is changed into the voice data which is a digital signal. This voice data is supplied to the feature-extraction section 52.

[0041] About the voice data from the AD translation section 51, for every suitable frame, the feature-extraction section 52 extracts feature parameters, such as a spectrum, and linear predictor coefficients, a cepstrum multiplier, a line spectrum pair, and supplies them to the characteristic quantity buffer 53 and the matching section 54. In the characteristic quantity buffer 53, the feature parameter from the feature-extraction section 52 is stored temporarily.

[0042] The matching section 54 recognizes the voice (input voice) inputted into the microphone 15, referring to the sound model database 55, the dictionary database 56, and the syntax database 57 if needed based on the feature parameter from the feature-extraction section 52, or the feature parameter memorized by the characteristic quantity buffer 53.

[0043] That is, the sound model database 55 has memorized the sound model showing the acoustical descriptions, such as each phoneme in the audio language which carries out speech recognition, and syllable. Here, as a sound model, HMM (Hidden Markov Model) etc. can be used, for example. The dictionary database 56 has memorized the word dictionary in which the information about the pronunciation was described about each word for recognition. The syntax database 57 has memorized the syntax rule each word registered into the word dictionary of the dictionary database 56 described it to be how it was carrying out a chain (connected). Here, as syntax rule, a context free language (CFG) and the regulation based on a statistical word chain probability (N-gram) etc. can be used, for example.

[0044] By referring to the word dictionary of the dictionary database 56, the matching section 54 is connecting the sound model memorized by the sound model database 55, and constitutes the sound model (word model) of a word. furthermore, the word model which connected the matching section 54 by referring to the syntax rule memorized by the syntax database 57 in some word models, and was connected by making it such -- using -- a feature parameter -- being based -- for example, HMM -- the voice inputted into the microphone 15 is recognized by law etc. And the speech recognition result by the matching section 54 is outputted in a text etc.

[0045] In addition, when to process again for the inputted voice is required, the matching section 54 processes using the feature parameter memorized by the characteristic quantity buffer 53, and, thereby, needs to require utterance for the second time of a user.

[0046] Drawing 6 is drawing showing the internal configuration of image recognition section 31B. The image picturized by CCD16 is inputted into the AD translation section 61 of image recognition section 31B, is changed into digital image data, and is outputted to the feature-extraction section 62. The feature-extraction section 62 performs feature extractions, such as edge detection of an object, and concentration change of an image, from the inputted image data, and calculates characteristic quantity, such as a feature parameter or a feature vector.

[0047] The characteristic quantity extracted by the feature-extraction section 62 is outputted to the face detecting element 63. The face detecting element 63 detects a user's face from the inputted characteristic quantity, and outputs the detection result to the distance test section 64. The distance test section 64 measures the sense of a face while measuring distance with a user using the detection result outputted from the face detecting element 63. The measured measurement result is outputted to the action decision mechanism section 33.

[0048] In addition, the distance with a user can be measured from change of the magnitude of a face. For example, it is possible to carry out by using the approach currently indicated by "Nerual Network-Based Frace Detection Henry A.Rowley, Shumeet Baluja, and and Takeo Kanade IEEE Pattern Analysis and Machine Intellegence."

[0049] Moreover, in the gestalt of this operation, although explained measuring magnitude of a face using one image input, distance with a user may be measured by performing matching during two image inputs (stereo image). It is Possible to Carry Out by Using Approach Currently Indicated by "Section 3.3.1 Point

Pattern-Matching Image-Analysis Handbook Mikio Takagi and Akihisa Shimoda Editorial-Supervision University of Tokyo Press" about Extract of Three-Dimension Information from Stereo Image, for Example.

[0050] While the characteristic quantity extracted by the feature-extraction section 62 is outputted to the face detecting element 63, it is outputted also to the matching section 65. The matching section 65 outputs the recognition result obtained by comparing the inputted characteristic quantity with the pattern information memorized by the standard-pattern database 66 to the action decision mechanism section 33. The data memorized by the standard-pattern database 66 are data in which the image data of gesture and the description of a pattern of operation are shown. in addition, recognition Robotics Society of Japan of a sensibility expression according to "gesture for example about gesture recognition -- it is possible to use the approach currently indicated by Vol.17 NO.7 933 page thru/or 936 pages, and 1999."

[0051] Thus, the recognition result outputted from speech recognition section 31A and the recognition result (measurement result) outputted from image recognition section 31B are inputted into the action decision mechanism section 33. Drawing 7 is drawing showing the internal configuration of the action decision mechanism section 33. The recognition result of the voice outputted from speech recognition section 31A is inputted into the text analysis section 71 of the action decision mechanism section 33. The text analysis section 71 extracts language information, such as information on a word, and information on functor, by analyzing the speech recognition result inputted based on the data memorized by the dictionary database 72 and the syntax database 73 for analysis for morphological analysis, syntax analysis, etc. Moreover, the semantics [an input] of voice utterance, an intention, etc. are extracted based on the contents described by the dictionary.

[0052] That is, information, such as part-of-speech information required for the dictionary database 72 in order to apply the notation and the syntax for analysis of a word, the semantic information according to individual of a word, etc. are memorized, and the data which described the constraint about a word chain based on the information on each word memorized by the dictionary database 72 are memorized by the syntax database 73 for analysis. The text analysis section 71 analyzes the inputted text data of a speech recognition result using these data.

[0053] The data memorized by the syntax database 73 for analysis are data required for text analysis using a language theory including semantics, such as HPSG, etc., when including even a regular grammar, a context free language, statistical word chain establishment, and semantic analysis.

[0054] The analysis result outputted from the text analysis section 71 is outputted to the keyword extraction section 74. From the inputted analysis result, with reference to the data memorized by the keyword database 75, the keyword extraction section 74 extracts the intention which the user uttered, and outputs the extract result to the table reference section 76 of operation. In addition, the data of language in which an intention of users, such as an admiration expression and an instruction, is shown are held as a keyword used for the keyword database 75 in the case of keyword spotting. Specifically, it is held as the expression which serves as an index of speech information in the latter table reference section 76 of operation, and data whose word corresponding to it is a keyword.

[0055] The table reference section 76 of operation is determined by referring to the table memorized by the table storage section 77 of operation and the classification table storage section 78 of operation, respectively in the actuation for which it opts by the extract result outputted from the keyword extraction section 74, and the recognition result outputted from image recognition section 31B. Here, the table memorized by the table storage section 77 of operation is explained. Drawing 8 is drawing showing an example of the table of operation memorized by the table storage section 77 of operation.

[0056] As a recognition result of an image, "beckoning", "handshaking" "which points at a finger", and when ["which shake a hand"] there "is" no recognition result of an image, it is classified. It is divided, when the measurement result of distance with a user is needed as attendant circumstances with these classifications, and when that is not right. Furthermore, actuation is determined by the audio recognition result.

[0057] For example, when it is "beckoning" as a result the user has recognized the image to be, the

information on being separated from where how many it is of the user first etc., i.e., a measurement result, is needed. and -- a user -- beckoning -- **** -- even if -- the -- the time -- utterance -- " -- here -- come -- " -- it is -- if -- " -- a user -- approaching -- " -- ** -- saying -- actuation -- determining -- having -- although -- " -- being suitable -- it can go -- " -- etc. -- it is -- if -- " -- a user -- from -- separating -- " -- actuation -- determining -- having . In addition, although mentioned later for details, even when "coming here" and a user speak, it does not necessarily surely opt for actuation of approaching a user.

[0058] Thus, a table of operation is a table described that a user's gesture (recognition result of an image), a user's utterance (audio recognition result), and three information further of distance (measurement result) with a user by the situation opt for one actuation.

[0059] Drawing 9 is drawing showing an example of the classification table of operation memorized by the classification table storage section 78 of operation. A classification table of operation classifies the actuation in a table of operation. Actuation of a table of operation can be classified into four kinds as shown in the classification table of operation shown in drawing 9 . That is, they are "robot location relative actuation", "user location relative actuation", an "absolute position action", and "others."

[0060] "Robot location relative actuation" is actuation as which the direction of operation and distance are determined in a robot's current position, for example, since a user's right-hand side turns into a robot's 1 left-hand side when a user can go to "right and speaks with ", and the robot 1 and the user have met, a robot 1 performs as a result actuation it is supposed that is moved to the left.

[0061] "User location relative actuation" is actuation as which the direction of operation and distance are determined in a user's current position, for example, when a user speaks with "come here", a robot 1 judges how much it should move, although it goes till the place 80cm before a user, and performs actuation of moving according to the decision to it.

[0062] An "absolute position action" is actuation which does not need the current position of a robot 1 or a user, for example, by a user's utterance inviting to "east, when it is ", since a robot 1 is the direction of the meaning determined nothing [say / east] with regards to a self location and a user's location, it performs actuation of moving in the direction.

[0063] "Others" are actuation which does not need the information on a direction or distance, for example, are that a robot 1 utters voice etc.

[0064] Next, the method of the decision of a robot 1 of operation made in a robot 1 is explained. As mentioned above, actuation of a robot 1 is determined by a user's voice and actuation. Then, the actuation which recognizes a user's voice is first explained with reference to the flow chart of drawing 10 . As for a user's voice incorporated with the microphone 15, processing of speech recognition is performed by speech recognition section 31A in step S1.

[0065] The recognition result outputted from speech recognition section 31A is inputted into the text analysis section 71 of the action decision mechanism section 33 in step S2, and text analysis is performed. And in step S3, keyword matching is performed by the keyword extraction section 74 using the result of the analysis. Consequently, it is judged in step S4 whether the keyword was extracted or not. In step S4, when it is judged that the keyword was extracted, it progresses to step S5.

[0066] Let the extracted keyword be language information in step S5. On the other hand, when it is judged in step S4 that a keyword is not extracted, it progresses to step S6 and let information that there is no keyword be language information. Termination of processing of step S5 or step S6 outputs language information to the table reference section 76 of operation in step S7. Such processing is repeatedly performed, while the robot 1 is operating.

[0067] While such speech recognition processing is performed, processing about a user's image is also performed. The image processing performed in a robot 1 is explained with reference to the flow chart of drawing 11 . As for the image picturized by CCD16, characteristic quantity is extracted by the feature-extraction section 62 of image recognition section 31B in step S11. The recognition result is used and it is judged in step S12 whether there is any gesture registered. That is, it judges whether the matching

section 65 has a match in the pattern information on the gesture memorized by the standard-pattern database 66 using the characteristic quantity outputted by the feature-extraction section 62. When it is judged by such decision that there is gesture, it progresses to step S13.

[0068] In step S13, it is judged whether the gesture judged to be gesture is a thing with incidental information. As gesture with incidental information, it is a case so that a user may require a direction predetermined with a finger very, and, in such a case, the information on an object that it is located in the direction in which the finger is present very turns into incidental information, for example. In step S13, when it is judged that it is gesture with incidental information, detection of the incidental information is performed in step S14. In step S14, after detection of incidental information is ended, it progresses to step S15.

[0069] When it is judged that there is no gesture registered in step S12 on the other hand, or also when it is judged in step S13 that there is no incidental information, it progresses to processing of step S15. In step S15, performance information is outputted to the table reference section 76 of operation.

[0070] When it progresses to processing of step S12 to the step S15, as performance information, they are the information that there is no gesture, and the information that there will be no information which opts for actuation as a recognition result of an image if it puts in another way. When it progresses to processing of step S13 to the step S15, as performance information, it is only the information about gesture. When it progresses to processing of step S14 to the step S15, as performance information, they are the information about gesture, and incidental information.

[0071] Such image recognition processing is repeatedly performed, while the robot 1 is operating. In addition, the measurement result outputted as incidental information on step S13 as a result of processing by the face detecting element 63 and the distance test section 64 is also included if needed.

[0072] Thus, the table reference section 76 of the action decision mechanism section 33 of operation opts for a robot's 1 action using the language information as a speech recognition result, and the performance information as an image recognition result. With reference to the flow chart of drawing 12 , actuation of the table reference section 76 of operation is explained. In step S21, performance information is inputted for language information from image recognition section 31B from the keyword extraction section 74, respectively. In step S22, actuation is determined as a meaning based on the language information and performance information which were inputted with reference to the table of operation memorized by the table storage section 77 of operation and the classification table of operation memorized by the classification table storage section 77 of operation.

[0073] Here, the actuation for which it opts is explained. although it opts for actuation based on the table of operation shown in drawing 8 , the recognition result (performance information) of an image is "beckoning", and when an audio recognition result (language information) is the "** here **", as actuation, actuation of three kinds of the user approaching a user who separates from a user being disregarded is set up, for example. always performing the same actuation, although actuation of approaching a user should be chosen if it is usual, and "it is beckoned" and called the "** here **" -- if -- it may get bored.

[0074] Then, even when a user does the same gesture and does the same utterance, it is made to make different actuation perform more each time. Then, determining [to which it is set] whether it is decided among three kinds of actuation which actuation it will be by feeling in case [that] the keyword which is determined in order, which is determined at random and which is determined by the probability value determines is considered.

[0075] When a probability value determines, the rate of whether it is decided which actuation it will be is beforehand determined like 50% "which approaches", 30% "to leave", and 20% "disregarded."

[0076] When a keyword determines, it is possible to carry out with the combination of current actuation, utterance, and the previous actuation and utterance. For example, when a user needs to strike a hand as pre-actuation and needs to beckon as current actuation, and it sets up so that actuation of surely approaching a user may be chosen, when come here is said, and knock as pre- actuation, beckon as current actuation, and come here is said, it sets up so that actuation of separating from a user may be chosen.

[0077] Thus, you may make it the combination of pre- actuation, utterance, and current actuation and utterance determine actuation.

[0078] If beckon and come here is said when approaching a user and sensing the resentment, if beckon and come here is said when the feeling at that time determines, and sensing fear by the feeling at that time with reference to the information on feeling / instinct model section 32, it is also possible to make it say that a user is disregarded.

[0079] Thus, the table reference section 76 of operation opts for actuation with reference to a table of operation based on language information and performance information. And the actuation for which it opted is outputted to the posture transition device section 34 in step S23 (drawing 12), and a robot 1 performs actuation for which it opted by performing processing predetermined in the part after it.

[0080] Although the direction which a user shows is detected from the direction which a user's finger puts and the object which exists in the direction was detected as incidental information in the gestalt of operation mentioned above, a direction is detected from the direction of a user's face, the direction which the eye has turned to, the direction which a jaw puts, and you may make it detect incidental information.

[0081] Moreover, the sign which shows the intimidation and the piece sign which close O.K. sign, BATSUMAKU, the round-head mark, safe, and the lug other than the gestalt of operation mentioned above (it is not audible), ******** (a palm is swayed horizontally), and money, the wish, a prayer, a hand, banzai, etc. become possible [using] by memorizing the information on the gesture generally used in the standard-pattern database 66.

[0082] When recognizing that the user spoke, the utterance itself is ambiguous (it does not speak clearly), and it may incorrect-recognize as speech recognition. For example, although a user takes "apple and speaks with ", since the utterance is not uttered clearly, as a result of incorrect recognition of speech recognition section 31A, "parakeet is taken and it may be recognized as "etc. In such a case, by using image data explains about how to identify whether it is an apple and whether it is a parakeet with reference to the flow chart of drawing 13 .

[0083] In step S31, if a user speaks, the voice will be incorporated by the robot 1 with a microphone 15, and will be inputted into speech recognition section 31A. Speech recognition section 31A recognizes the inputted sound signal in step S32. And two or more candidates judged that the user probably spoke as the result are mentioned. Processing of step S33 is performed to the probable candidate of the 1st place, and the candidate of the 2nd place among those candidates.

[0084] In step S33, it is judged whether the difference of the score value of the candidate of the 1st place and the candidate of the 2nd place is less than a predetermined threshold. Consequently, if it puts in another way and the candidate of the 1st place will be judged to be satisfactory as a recognition result since the difference of the score value of the candidate of the 1st place and the score value of the candidate of the 2nd place is separated when it is judged that it is not less than a predetermined threshold, it will progress to step S37, the recognition result will be decided as a speech recognition result, and it will be used.

[0085] If it is judged that the 1st candidate may be incorrect recognition if the difference of the score value of the candidate of the 1st place and the score value of the candidate of the 2nd place is judged to be less than a threshold in step S33 and it will put in another way on the other hand, it will progress to step S34 and let two or more candidates with an expensive score be processing objects as a recognition result. Image recognition is performed in step S36. The image picturized when it was the order at the time of the image picturized when utterance of the user who is the processing object of speech recognition was carried out, or utterance being carried out is a processing-object image of the image recognition in step S35.

[0086] In step S36, the complement as a result of speech recognition is performed using the result of the image recognition in step S35.

[0087] For example, as mentioned above, when a user takes "apple and speaks with ", as the recognition result, the candidate of the 1st place takes "apple, and it is ", and the candidate of the 2nd place takes "parakeet and suppose that it was ". Furthermore, when these candidates of the 1st place and the candidate of

the 2nd place are less than predetermined thresholds, which candidate cannot judge in that of the right. Then, the picturized image is then recognized, for example, the parakeet which is the candidate of the 1st place when it is judged that the apple is picturized in the image and which "is the candidate of the 2nd place when an apple is taken, it judges that "is as a result of right recognition and it is judged that the parakeet is picturized in an image" is taken, and it is judged that "is as a result of right recognition.

[0088] Thus, a complement of the result of speech recognition decides the complemented speech recognition result as a speech recognition result in step S37. Thus, when ambiguity is contained in a recognition result, it becomes possible by using image information to perform speech recognition more certainly.

[0089] In addition, in the explanation mentioned above, although only the difference of the score value of the candidate of the 1st place and the candidate of the 2nd place was compared, approaches, such as taking the difference of the candidate of the 10th place from the candidate of the 1st place, may be used.

[0090] by the way, the time of User A and User B talking -- User A -- " -- suppose that tried to be fastidious and it spoke with ". this utterance -- receiving -- User B -- " -- is it in it? It speaks with ". Such a conversation is a conversation exchanged well in every day. That is, a demonstrative pronoun changes with situations at that time as it is "it", if it is "this" for User A and takes to User B also to the same object.

[0091] Such a thing is being able to say, when the robot's 1 is talking with the user, therefore a robot 1 needs to recognize clearly to what the user is pointing. Processing of the robot 1 when recognizing the object which a demonstrative pronoun shows is explained with reference to the flow chart of drawing 14 . In step S41, a user speaks and speech recognition is performed in step S42 about the utterance.

[0092] In step S43, it is judged using the result of speech recognition whether a demonstrative pronoun is in a user's utterance. If it is judged that there is no demonstrative pronoun, the result of the speech recognition will be decided as a speech recognition result in step S46.

[0093] When it is judged that a demonstrative pronoun is in the inside which the user uttered in step S43 on the other hand, it progresses to step S44 and image recognition is performed. The image set as the object of image recognition is an image in the time of picturizing the image which judges the image picturized when the user spoke, or the direction which a user puts with a finger etc., and exists in the direction.

[0094] In step S44, if the image recognition of the picturized image is performed, in step S45, the complement of a demonstrative pronoun will be performed using the recognition result (image information). Here, a concrete example is given and explained. a user -- " -- suppose that it swerved and spoke to "and a robot 1. A user takes gesture, such as pointing to the object corresponding to "it" with a finger, in that case.

[0095] A robot 1 receives the utterance, and performs speech recognition in step S42, consequently judges that "it" which is a demonstrative pronoun is included. Moreover, it is judged that the user has taken the gesture of pointing to a direction predetermined with a finger, from the image picturized when the user spoke.

[0096] In step S44, a user judges the direction where it pointed with "it", and picturizes the image of the direction, and a robot 1 performs image recognition to the picturized image. If recognized as an object as a result of the image recognition (for example, a newspaper), the object which the demonstrative pronoun "it" shows will be complemented as it is a "newspaper." Thus, in step S45, if a demonstrative pronoun is complemented from image information, it will progress to step S46 and the complemented speech recognition result will be decided as a speech recognition result.

[0097] Thus, it becomes possible by using image information to recognize certainly the object which a demonstrative pronoun shows.

[0098] When a robot 1 picturizes an image, in the picturized image, two or more bodies are included in many cases. The body to which conversation carries out an object and the user is pointing among such two or more bodies is explained with reference to the flow chart of drawing 15 about the processing which recognizes what it is. In step S51, a user's gesture picturized by CCD16 is inputted into a robot 1 as an image.

[0099] When it is said by inputting gesture that the gesture points to a predetermined direction for example, in order to detect incidental information, it is necessary to recognize the image of the direction to which a user points. Then, the image of the direction to which a user points is picturized and image recognition

processing by image recognition section 31B is performed in step S52 about the image. The recognition result is used and it is judged in step S53 whether two or more objects exist in an image. In step S53, if it is judged that two or more objects do not exist, it will put in another way and it will be judged that only one exists as an object, it will progress to step S56 and the image recognition result of the object will be outputted as an image recognition result.

[0100] On the other hand, when an object is judged that there are more than one in step S53, it progresses to step S54 and speech recognition is performed. The voice set as the object of speech recognition is the voice incorporated when the user performed gesture. The result (speech information) of the speech recognition in step S54 is used, and the complement of an image recognition result is performed in step S55. Here, an example is given and explained.

[0101] While a user does the gesture of pointing to a predetermined direction, suppose that "ball was taken and it spoke with ". First, a robot 1 recognizes a user's gesture and recognizes it as the gesture being gesture indicating a predetermined direction. And the image of the direction to which it points is picturized, and the object in an image is recognized. Consequently, if it is judged that two or more objects exist, voice which the user uttered to gesture and coincidence will be recognized.

[0102] If "ball is taken and it is recognized as "as a result of the speech recognition, the "ball" of them will be judged to be the object which the user is considering as the request also in two or more objects in an image. That is, an image recognition result is complemented from speech information. Thus, if an image recognition result is complemented from speech information, the image recognition result progressed and complemented will be outputted to step S56 as an image recognition result.

[0103] Thus, it becomes possible using speech information to acquire image information with a more high precision by complementing the ambiguous part of image information.

[0104] By the way, the robot which acts and takes action only by image information that the robot which takes action only by speech information advances in the direction in which there is a user's voice performs action called ** in the direction in which a user is settled in the image currently picturized, for example. However, as mentioned above, the robot 1 which applied this invention judges the actuation for which the user is asking combining speech information and image information, and actually shifts to action. Then, as it already explained that a robot's 1 action was classified, and shown in a classification table of operation as shown in drawing 9 , it can classify.

[0105] That is, it opts for actuation by recognizing voice and grasping the location of user and robot 1 self from image information. When a user speaks with "come here", first, speech recognition of the utterance is carried out, next, specifically, a user's location is recognized from image information. And when it opts for actuation of progressing in the direction of a user, the purpose location in which direction it progresses is determined in the distance of which.

[0106] For example, as shown in drawing 16 , as a purpose location, it is set up with the place 80cm before a user. The distance between selves is measured with a user using the characteristic quantity from which the feature-extraction section 62 (drawing 6) of image recognition section 31B extracted this based on the magnitude of a user's face which the face detecting element 63 has recognized a user's face, and has been recognized by the distance test section 64. And it is determined which should progress in order to move to 80cm of a user's this side using the measured distance.

[0107] Thus, it becomes possible by measuring a user's location and using a user's location according to actuation to make actuation to a user's gesture more exact.

[0108] In the gestalt of operation mentioned above, although the language which actually spoke as a user's voice was recognized, a user can opt for actuation of a robot 1 as voice using the sound (rhythm) made with the handclap, and a user's footstep (sound).

[0109] When using a user's rhythm and sound (a thing including these is hereafter expressed as voice suitably), the configuration of speech recognition section 31A turns into a configuration as shown in drawing 17 . That is, the voice incorporated with the microphone 15 is inputted into the AD translation section 51, is

changed into digital data and is further inputted into a rhythm / sound recognition section 81. A rhythm / sound recognition section 81 acquires the information about a rhythm or sound.

[0110] The recognition result acquired by a rhythm / sound recognition section 81 is outputted to the action decision mechanism section 33. In addition, to drawing 17 , the part which recognizes a user's utterance, i.e., the part shown in drawing 5 , is omitted and described. Therefore, the digital sound signal outputted from the AD translation section 51 is outputted also to a rhythm / sound recognition section 81 (drawing 17) while it is outputted to the feature-extraction section 52 (drawing 5).

[0111] Furthermore, although the recognition result outputted from a rhythm / sound recognition section 81 is outputted to the action decision mechanism section 33, it is directly inputted into the table reference section 76 of operation rather than is inputted into the text analysis section 71 (drawing 7) of the account posterior part 33 of action decision.

[0112] Here, the recognition approach of the rhythm which a rhythm / sound recognition section 81 performs is explained. A rhythm is detected using approaches, such as beat (beat) detection of a rhythm / sound recognition section 81 percussion-instrument sound (the sound by a user's handclap is included), or beat detection by code (chord) change. Consequently, the beat was detected when or how many beat child and the detection result of several beats per minute is outputted.

[0113] The sound-source separation system for "percussion instrument sound about the approach of rhythm detection, Goto true **, Yoichi Muraoka work, the Institute of Electronics, Information and Communication Engineers paper magazine, J77-DII, "NO.5,901 thru/or 911 pages, and 1994", the real-time beat tracking system for an acoustic signal, It is possible to use the approach currently indicated by Goto true **, Yoichi Muraoka work, the Institute of Electronics, Information and Communication Engineers paper magazine, J81-DII, NO.2,227 or 237 pages, 1998", etc.

[0114] Using the recognition result about the rhythm outputted from a rhythm / sound recognition section 81, the case where it dances as actuation for which the action decision mechanism section 33 (table reference section 76 of operation) opts is mentioned as an example, and is explained here. The table of operation as shown in drawing 18 is memorized by the table storage section 77 of operation. for example, the recognition results about a rhythm are 0 thru/or 60 beats in 1 minute, in the case of two rhythm, dance A is chosen, and in 1 minute, when it is 0 thru/or 60 beats and is not two rhythm, three rhythm, or four rhythm, either, dance A chooses -- having -- ****** -- like -- ** for 1 minute, and several beats -- ** -- the type of a dance is chosen as a meaning by the information to say.

[0115] Thus, a robot 1 is controlled by performing processing predetermined in the part after the action decision mechanism section 33 so that actuation for which it opted by referring to the table of operation where the table reference section 76 of operation is memorized by the table storage section 77 of operation is performed.

[0116] Although the information about a rhythm was acquired with voice, you may make it acquired by gesture in the explanation mentioned above. When the information about a rhythm is acquired by gesture, the configuration of image recognition section 31B is good with a configuration as shown in drawing 6 . It is possible to use recognition of the sensibility expression by "gesture, Seiji Iguchi work, and the approach currently indicated by 17 Robotics Society of Japan No. 7" as an approach of acquiring the information about a rhythm by gesture.

[0117] Of course, you may make it acquire the information about a rhythm from both voice and gesture.

[0118] Next, the case where sound opts for actuation of a robot 1 is explained. It is shown whether they are sounds which who or what emitted, such as a sound in which the favorite person emitted a scream etc. and what kinds of sounds [a footstep and] they were as a sound recognized by a rhythm / sound recognition section 81 again, for example, a sound which the disagreeable person emitted, and a sound which the vehicle emitted.

[0119] The result recognized in a rhythm / sound recognition section 81 is outputted to the table reference section 76 of operation. The table reference section 76 of operation opts for the actuation corresponding to

the recognition result about the inputted sound with reference to the table of operation memorized by the table storage section 77 of operation. An example of the table of operation about the sound memorized by the table storage section 77 of operation is shown in drawing 19 .

[0120] as a table of operation shown in drawing 19 , for example as a recognition result of acoustic, a footstep is detected, and if the footstep is judged to be a favorite person's footstep, actuation that joy approaches will choose -- having -- ****** -- it is the table where action is determined as a meaning according to a sound phenomenon like. A robot 1 judges the information of a favorite person and a disagreeable person from the conversation exchanged between a user and a robot 1, a user's attitude, etc., and you may make it memorize it as information.

[0121] Moreover, you may make it use not only sound but image information. That is, although it is also possible to judge whether someone came from the footstep when a footstep can be heard, it is picturized as an image, and the man judges whether you are a favorite person and whether you are a disagreeable person, and may be made judge who it is and to opt for actuation from the recognized result.

[0122] As mentioned above, by combining speech information and image information, it becomes possible to make various actuation give a robot 1, and it becomes possible further by using mutual information in the phase of recognition of the voice in the decision of operation, and an image to perform institutional high recognition processing more.

[0123] Although a series of processings mentioned above can also be performed by hardware, they can also be performed with software. When performing a series of processings with software, the program which constitutes the software is installed in a general-purpose personal computer etc. from a record medium possible [performing various kinds of functions] by installing the computer built into the hardware of dedication, or various kinds of programs.

[0124] As shown in drawing 20 , this record medium is distributed apart from a computer in order to provide a user with a program. The magnetic disk 131 (a floppy disk is included) with which the program is recorded, an optical disk 132 (CD-ROM (Compact Disk-Read Only Memory) --) DVD (Digital Versatile Disk) is included. It is not only constituted by the package media which consist of a magneto-optic disk 133 (MD (Mini-Disk) is included) or semiconductor memory 134, but It consists of hard disks with which ROM112 with which a user is provided in the condition of having been beforehand included in the computer, and the program is remembered to be, and the storage section 118 are contained.

[0125] In addition, in this specification, even if the processing serially performed according to the sequence that the step which describes the program offered by the medium was indicated is not of course necessarily processed serially, it is a juxtaposition thing also including the processing performed according to an individual.

[0126] Moreover, in this specification, a system expresses the whole equipment constituted by two or more equipments.

[0127]

[Effect of the Invention] Since voice is recognized, an image is recognized and it opted for actuation of a robot among the audio recognition result or the recognition result of an image at least using one side according to the information processor according to claim 1, the information processing approach according to claim 10, and the record medium according to claim 11 like the above, it becomes possible to perform speech recognition and image recognition with a more high system.

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]

[Drawing 1] It is the perspective view showing the example of an appearance configuration of the gestalt of 1 operation of the robot which applied this invention.

[Drawing 2] It is the block diagram showing the example of an internal configuration of the robot of drawing 1.

[Drawing 3] It is the block diagram showing the example of a functional configuration of the controller 10 of drawing 2.

[Drawing 4] It is drawing showing the example of a functional configuration about the part which recognizes voice and an image and opts for action.

[Drawing 5] It is the block diagram showing the internal configuration of speech recognition section 31A.

[Drawing 6] It is the block diagram showing the internal configuration of image recognition section 31B.

[Drawing 7] It is the block diagram showing the internal configuration of the action decision mechanism section 33.

[Drawing 8] It is drawing explaining the table of operation memorized by the table storage section 77 of operation.

[Drawing 9] It is drawing explaining the classification table of operation memorized by the classification table storage section 78 of operation.

[Drawing 10] It is a flow chart explaining speech recognition processing.

[Drawing 11] It is a flow chart explaining image recognition processing.

[Drawing 12] It is a flow chart explaining decision processing of operation.

[Drawing 13] It is the flow chart which explains the processing in the case of outputting a recognition result using speech information and image information.

[Drawing 14] It is the flow chart which explains other processings in the case of outputting a recognition result using speech information and image information.

[Drawing 15] It is the flow chart which explains the processing of further others in the case of outputting a recognition result using speech information and image information.

[Drawing 16] It is drawing explaining the physical relationship of a user and a robot 1.

[Drawing 17] It is drawing showing other configurations of speech recognition section 31A.

[Drawing 18] It is drawing explaining other tables of operation memorized by the table storage section 77 of operation.

[Drawing 19] It is drawing explaining the table of further others of operation memorized by the table storage section 77 of operation.

[Drawing 20] It is drawing explaining a medium.

[Description of Notations]

10 Controller 10A CPU, 10B Memory 15 A microphone and 16 CCD 17 Touch sensor 18 loudspeaker 19 The communications department, 31 Sensor input-process section 31A Speech recognition section 31B Image recognition section 31C Pressure processing section 32 Feeling / instinct model section 33 The action decision mechanism section and 34 posture transition device section 35 Controlling mechanism section 36 Speech synthesis section

---

CLAIMS

[Claim(s)]

[Claim 1] The information processor characterized by including a speech recognition means to recognize voice, an image recognition means to recognize an image, and a decision means to opt for actuation of said robot at least using one side among the recognition result by said speech recognition means, or the recognition result by said image recognition means, in the information processor built in a robot.

[Claim 2] The information processor according to claim 1 characterized by including further a maintenance means to hold the table on which the relation of said robot of operation determined as a meaning by the recognition result by said speech recognition means, the recognition results by said image recognition means, and those recognition

results was described.

[Claim 3] Said decision means is an information processor according to claim 1 characterized by deciding that it will be a meaning and opting for actuation of said robot using the determined recognition result using the recognition result by said image recognition means when the recognition result by said voice means is not determined as a meaning.

[Claim 4] Said decision means is an information processor according to claim 1 characterized by opting for actuation of said robot using the recognition result determined by the meaning in said image which said image recognition means recognizes using the recognition result by said speech recognition means when two or more objects exist.

[Claim 5] Said image recognition means is an information processor according to claim 1 characterized by recognizing the image which detects the direction to which the part beforehand set up among a user's finger, the face, the eye, and the jaw points, and is located in the direction.

[Claim 6] Said image recognition means is an information processor according to claim 1 characterized by detecting the gesture memorized by said storage means and making the detected result into a recognition result by recognizing said user's image, including further a storage means to memorize the data about the gesture which a user performs.

[Claim 7] It is the information processor according to claim 1 characterized by also using the measurement result by said measurement means for said decision means, and opting for actuation of said robot from the magnitude of said user's face detected by detection means to detect a user's face, and said detection means, including further a measurement means to measure the distance between said user and self.

[Claim 8] Said speech recognition means is an information processor according to claim 1 characterized by detecting the rhythm contained in an environmental sound and considering as a recognition result.

[Claim 9] Said speech recognition means is an information processor according to claim 1 characterized by detecting a sound phenomenon and considering as a recognition result from an environmental sound.

[Claim 10] The information-processing approach characterized by to include the decision step which opts for actuation of said robot in the information-processing approach of the information processor built in a robot at least using one side among the recognition result by processing of the speech recognition step which recognizes voice, the image recognition step which recognizes an image, and said speech recognition step, or the recognition result by processing of said image recognition step.

[Claim 11] The record medium with which the program which the computer characterized by to be included the decision step which opts for actuation of said robot at least using one side among the recognition result by processing of the speech-recognition step which is a program for information processing of the information processor built in a robot, and recognizes voice, the image-recognition step which recognizes an image, and said speech-recognition step, or the recognition result by processing of said image-recognition step can read is recorded.

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2001－188555
（P2001－188555A）
(43)公開日　平成13年７月10日(2001.7.10)

(51)Int.Cl.'	識別記号	FI		テーマコード'（参考）
Ｇ１０Ｌ　15/00		Ｂ２５Ｊ　13/00	Ｚ	３Ｆ０５９
Ｂ２５Ｊ　13/00		Ｇ１０Ｌ　3/00	５５１Ｈ	５Ｂ０５７
Ｇ０６Ｔ　1/00		Ｇ０６Ｆ　15/62	３８０	５Ｄ０１５
7/20		15/70	４１０	５Ｌ０９６
Ｇ１０Ｌ　15/24		Ｇ１０Ｌ　3/00	５７１Ｑ	９Ａ００１

審査請求　未請求　請求項の数11　ＯＬ　（全 16 頁）

(21)出願番号　特願平11－375773

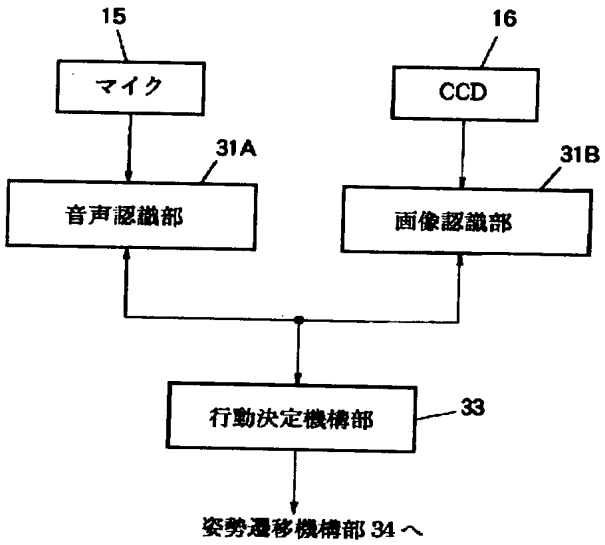
(22)出願日　平成11年12月28日(1999.12.28)

(71)出願人　000002185
ソニー株式会社
東京都品川区北品川６丁目７番35号
(72)発明者　山下　潤一
東京都品川区北品川６丁目７番35号　ソニー株式会社内
(72)発明者　小川　浩明
東京都品川区北品川６丁目７番35号　ソニー株式会社内
(74)代理人　100082131
弁理士　稲本　義雄

最終頁に続く

(54)【発明の名称】　情報処理装置および方法、並びに記録媒体

(57)【要約】
【課題】　バラエティに富んだ動作を行うロボットを提供する。
【解決手段】　マイク１５により取り込まれたユーザの音声は、音声認識部３１Ａにより音声認識される。ＣＣＤ１６により撮像されたユーザのジェスチャは、画像認識部３１Ｂにより認識される。行動決定機構部３３は、音声認識部３１Ａから出力される音声情報と、画像認識部３１Ｂから出力される画像情報とを用いて、ロボットの動作を決定する。



【特許請求の範囲】

【請求項１】　ロボットに内蔵される情報処理装置において、

音声を認識する音声認識手段と、

画像を認識する画像認識手段と、

前記音声認識手段による認識結果、または、前記画像認識手段による認識結果のうち、少なくとも一方を用いて、前記ロボットの動作を決定する決定手段とを含むことを特徴とする情報処理装置。

【請求項２】　前記音声認識手段による認識結果、前記画像認識手段による認識結果、および、それらの認識結果により一意に決定される前記ロボットの動作の関係について記述されたテーブルを保持する保持手段をさらに含むことを特徴とする請求項１に記載の情報処理装置。

【請求項３】　前記決定手段は、前記音声手段による認識結果が、一意に決定されない場合、前記画像認識手段による認識結果を用いて、一意に決定し、その決定された認識結果を用いて、前記ロボットの動作を決定することを特徴とする請求項１に記載の情報処理装置。

【請求項４】　前記決定手段は、前記画像認識手段が認識する前記画像内に、複数の対象物が存在する場合、前記音声認識手段による認識結果を用いて、一意に決定される認識結果を用いて、前記ロボットの動作を決定することを特徴とする請求項１に記載の情報処理装置。

【請求項５】　前記画像認識手段は、ユーザの指、顔、目、あごのうち、予め設定された部分が指し示す方向を検出し、その方向に位置する画像を認識することを特徴とする請求項１に記載の情報処理装置。

【請求項６】　ユーザの行うジェスチャに関するデータを記憶する記憶手段をさらに含み、

前記画像認識手段は、前記ユーザの画像を認識することにより、前記記憶手段に記憶されているジェスチャを検出し、その検出された結果を認識結果とすることを特徴とする請求項１に記載の情報処理装置。

【請求項７】　ユーザの顔を検出する検出手段と、

前記検出手段により検出された前記ユーザの顔の大きさから、前記ユーザと自己との間の距離を測定する測定手段とをさらに含み、

前記決定手段は、前記測定手段による測定結果も用いて、前記ロボットの動作を決定することを特徴とする請求項１に記載の情報処理装置。

【請求項８】　前記音声認識手段は、環境音に含まれるリズムを検出し、認識結果とすることを特徴とする請求項１に記載の情報処理装置。

【請求項９】　前記音声認識手段は、環境音から、音響現象を検出し、認識結果とすることを特徴とする請求項１に記載の情報処理装置。

【請求項１０】　ロボットに内蔵される情報処理装置の情報処理方法において、

音声を認識する音声認識ステップと、

画像を認識する画像認識ステップと、

前記音声認識ステップの処理による認識結果、または、前記画像認識ステップの処理による認識結果のうち、少なくとも一方を用いて、前記ロボットの動作を決定する決定ステップとを含むことを特徴とする情報処理方法。

【請求項１１】　ロボットに内蔵される情報処理装置の情報処理用のプログラムであって、

音声を認識する音声認識ステップと、

画像を認識する画像認識ステップと、

前記音声認識ステップの処理による認識結果、または、前記画像認識ステップの処理による認識結果のうち、少なくとも一方を用いて、前記ロボットの動作を決定する決定ステップとを含むことを特徴とするコンピュータが読み取り可能なプログラムが記録されている記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は情報処理装置および方法、並びに記録媒体に関し、特に、音声情報と画像情報とを用いて、動作を決定するようなロボットに用いて好適な情報処理装置および方法、並びに記録媒体に関する。

【０００２】

【従来の技術】従来より、玩具等として、タッチスイッチが押圧操作されると、合成音を出力するロボット（ぬいぐるみ状のものを含む）や、ユーザの発話を認識し、応答文を返すという擬似的な会話を楽しめるロボットなどが数多く製品化されている。

【０００３】また、画像を撮像し、その撮像した画像を基に画像認識を行い、状況判断を行い、自律的に行動するロボットなども製品化されている。

【０００４】

【発明が解決しようとする課題】しかしながら、音声認識による認識結果は、例えば、ユーザの発話がはっきりと行われなかった場合など、誤認識をしてしまうといった課題があった。また、曖昧な言葉を含む発話、例えば、指示代名詞を含む発話の場合、指示代名詞が指示する対象物が認識できないときがあるという課題があった。

【０００５】また、上述したロボットは、音声または画像の一方の情報のみに依存して自律的な動作を行うものであり、音声と画像の情報を両方とも用いて自律的な動作を行う事は困難であった。

【０００６】本発明はこのような状況に鑑みてなされたものであり、音声と画像の両方の情報を用いることにより、より確実な音声認識を行い、バラエティに富んだ動作をユーザに提供することを目的とする。

【０００７】

【課題を解決するための手段】請求項１に記載の情報処理装置は、音声を認識する音声認識手段と、画像を認識する画像認識手段と、音声認識手段による認識結果、ま

たは、画像認識手段による認識結果のうち、少なくとも一方を用いて、ロボットの動作を決定する決定手段とを含むことを特徴とする。

【０００８】前記音声認識手段による認識結果、画像認識手段による認識結果、および、それらの認識結果により一意に決定されるロボットの動作の関係について記述されたテーブルを保持する保持手段をさらに含むようにすることができる。

【０００９】前記決定手段は、音声手段による認識結果が、一意に決定されない場合、画像認識手段による認識結果を用いて、一意に決定し、その決定された認識結果を用いて、ロボットの動作を決定するようにすることができる。

【００１０】前記決定手段は、画像認識手段が認識する画像内に、複数の対象物が存在する場合、音声認識手段による認識結果を用いて、一意に決定される認識結果を用いて、ロボットの動作を決定するようにすることができる。

【００１１】前記画像認識手段は、ユーザの指、顔、目、あごのうち、予め設定された部分が指し示す方向を検出し、その方向に位置する画像を認識するようにすることができる。

【００１２】ユーザの行うジェスチャに関するデータを記憶する記憶手段をさらに含み、画像認識手段は、ユーザの画像を認識することにより、記憶手段に記憶されているジェスチャを検出し、その検出された結果を認識結果とするようにすることができる。

【００１３】ユーザの顔を検出する検出手段と、検出手段により検出されたユーザの顔の大きさから、ユーザと自己との間の距離を測定する測定手段とをさらに含み、決定手段は、測定手段による測定結果も用いて、ロボットの動作を決定することを特徴とする。

【００１４】前記音声認識手段は、環境音に含まれるリズムを検出し、認識結果とするようにすることができる。

【００１５】前記音声認識手段は、環境音から、音響現象を検出し、認識結果とするようにすることができる。

【００１６】請求項１０に記載の情報処理方法は、音声を認識する音声認識ステップと、画像を認識する画像認識ステップと、音声認識ステップの処理による認識結果、または、画像認識ステップの処理による認識結果のうち、少なくとも一方を用いて、ロボットの動作を決定する決定ステップとを含むことを特徴とする。

【００１７】請求項１１に記載の記録媒体のプログラムは、音声を認識する音声認識ステップと、画像を認識する画像認識ステップと、音声認識ステップの処理による認識結果、または、画像認識ステップの処理による認識結果のうち、少なくとも一方を用いて、ロボットの動作を決定する決定ステップとを含むことを特徴とする。

【００１８】請求項１に記載の情報処理装置、請求項１０に記載の情報処理方法、および請求項１１に記載の記録媒体においては、音声が認識され、画像が認識され、音声の認識結果、または、画像の認識結果のうち、少なくとも一方を用いて、ロボットの動作が決定される。

【００１９】

【発明の実施の形態】図１は、本発明を適用したロボットの一実施の形態の外観構成例を示しており、図２は、その電気的構成例を示している。

【００２０】本実施の形態では、ロボットは、犬形状のものとされており、胴体部ユニット２の前後左右に、それぞれ脚部ユニット３Ａ，３Ｂ，３Ｃ，３Ｄが連結されるとともに、胴体部ユニット２の前端部と後端部に、それぞれ頭部ユニット４と尻尾部ユニット５が連結されることにより構成されている。

【００２１】尻尾部ユニット５は、胴体部ユニット２の上面に設けられたベース部５Ｂから、２自由度をもって湾曲または揺動自在に引き出されている。胴体部ユニット２には、ロボット全体の制御を行うコントローラ１０、ロボットの動力源となるバッテリ１１、並びにバッテリセンサ１２および熱センサ１３からなる内部センサ部１４などが収納されている。

【００２２】頭部ユニット４には、「耳」に相当するマイク（マイクロフォン）１５、「目」に相当するCCD(Charge Coupled Device)カメラ１６、触覚に相当するタッチセンサ１７、「口」に相当するスピーカ１８などが、それぞれ所定位置に配設されている。

【００２３】脚部ユニット３Ａ乃至３Ｄそれぞれの関節部分や、脚部ユニット３Ａ乃至３Ｄそれぞれと胴体部ユニット２の連結部分、頭部ユニット４と胴体部ユニット２の連結部分、並びに尻尾部ユニット５と胴体部ユニット２の連結部分などには、図２に示すように、それぞれアクチュエータ$３A A_1$乃至$３A A_x$、$３B A_1$乃至$３B A_x$、$３C A_1$乃至$３C A_x$、$３D A_1$乃至$３D A_x$、$４A_1$乃至$４A_t$、$５A_1$および$５A_2$が配設されており、これにより、各連結部分は、所定の自由度をもって回転することができるようになっている。

【００２４】頭部ユニット４におけるマイク１５は、ユーザからの発話を含む周囲の音声（音）を集音し、得られた音声信号を、コントローラ１０に送出する。CCDカメラ１６は、周囲の状況を撮像し、得られた画像信号を、コントローラ１０に送出する。

【００２５】タッチセンサ１７は、例えば、頭部ユニット４の上部に設けられており、ユーザからの「なでる」や「たたく」といった物理的な働きかけにより受けた圧力を検出し、その検出結果を圧力検出信号としてコントローラ１０に送出する。

【００２６】胴体部ユニット２におけるバッテリセンサ１２は、バッテリ１１の残量を検出し、その検出結果を、バッテリ残量検出信号としてコントローラ１０に送出する。熱センサ１３は、ロボット内部の熱を検出し、

その検出結果を、熱検出信号としてコントローラ１０に送出する。

【００２７】コントローラ１０は、CPU(Central Processing Unit)１０Ａやメモリ１０Ｂ等を内蔵しており、CPU１０Ａにおいて、メモリ１０Ｂに記憶された制御プログラムが実行されることにより、各種の処理を行う。即ち、コントローラ１０は、マイク１５や、CCDカメラ１６、タッチセンサ１７、バッテリセンサ１２、熱センサ１３から与えられる音声信号、画像信号、圧力検出信号、バッテリ残量検出信号、熱検出信号に基づいて、周囲の状況や、ユーザからの指令、ユーザからの働きかけなどの有無を判断する。

【００２８】さらに、コントローラ１０は、この判断結果等に基づいて、続く行動を決定し、その決定結果に基づいて、アクチュエータ３ＡＡ$_1$乃至３ＡＡ$_k$、３ＢＡ$_1$乃至３ＢＡ$_k$、３ＣＡ$_1$乃至３ＣＡ$_k$、３ＤＡ$_1$乃至３ＤＡ$_k$、４Ａ$_1$乃至４Ａ$_L$、５Ａ$_1$、５Ａ$_2$のうちの必要なものを駆動させ、これにより、頭部ユニット４を上下左右に振らせたり、尻尾部ユニット５を動かせたり、各脚部ユニット３Ａ乃至３Ｄを駆動して、ロボットを歩行させるなどの行動を行わせる。

【００２９】また、コントローラ１０は、必要に応じて、合成音を生成し、スピーカ１８に供給して出力させたり、ロボットの「目」の位置に設けられた図示しないLED (Light Emitting Diode) を点灯、消灯または点滅させる。

【００３０】以上のようにして、ロボットは、周囲の状況等に基づいて自律的に行動をとることができるようになっている。

【００３１】次に、図３は、図２のコントローラ１０の機能的構成例を示している。なお、図３に示す機能的構成は、CPU１０Ａが、メモリ１０Ｂに記憶された制御プログラムを実行することで実現されるようになっている。

【００３２】コントローラ１０は、特定の外部状態を認識するセンサ入力処理部３１、センサ入力処理部３１の認識結果等を累積して、感情および本能の状態を表現する感情／本能モデル部３２、センサ入力処理部３１の認識結果等に基づいて、続く行動を決定する行動決定機構部３３、行動決定機構部３３の決定結果に基づいて、実際にロボットに行動を起こさせる姿勢遷移機構部３４、各アクチュエータ３ＡＡ$_1$乃至５Ａ$_1$および５Ａ$_2$を駆動制御する制御機構部３５、合成音を生成する音声合成部３６、並びに音声合成部３６の出力を制御する音響処理部３７から構成されている。

【００３３】センサ入力処理部３１は、マイク１５や、CCDカメラ１６、タッチセンサ１７等から与えられる音声信号、画像信号、圧力検出信号等に基づいて、特定の外部状態や、ユーザからの特定の働きかけ、ユーザからの指示等を認識し、その認識結果を表す状態認識情報を、感情／本能モデル部３２および行動決定機構部３３に通知する。

【００３４】即ち、センサ入力処理部３１は、音声認識部３１Ａを有しており、音声認識部３１Ａは、行動決定機構部３３からの制御にしたがい、マイク１５から与えられる音声信号を用いて、音声認識を行う。そして、音声認識部３１Ａは、その音声認識結果としての、例えば、「歩け」、「伏せ」、「ボールを追いかけろ」等の指令その他を、状態認識情報として、感情／本能モデル部３２および行動決定機構部３３に通知する。

【００３５】また、センサ入力処理部３１は、画像認識部３１Ｂを有しており、画像認識部３１Ｂは、CCDカメラ１６から与えられる画像信号を用いて、画像認識処理を行う。そして、画像認識部３１Ｂは、その処理の結果、例えば、「赤い丸いもの」や、「地面に対して垂直なかつ所定高さ以上の平面」等を検出したときには、「ボールがある」や、「壁がある」等の画像認識結果を、状態認識情報として、感情／本能モデル部３２および行動決定機構部３３に通知する。また、ユーザが行うジェスチャの認識も行い、その認識結果を行動決定機構部３３に通知する。

【００３６】さらに、センサ入力処理部３１は、圧力処理部３１Ｃを有しており、圧力処理部３１Ｃは、タッチセンサ１７から与えられる圧力検出信号を処理する。そして、圧力処理部３１Ｃは、その処理の結果、所定の閾値以上で、かつ短時間の圧力を検出したときには、「たたかれた（しかられた）」と認識し、所定の閾値未満で、かつ長時間の圧力を検出したときには、「なでられた（ほめられた）」と認識して、その認識結果を、状態認識情報として、感情／本能モデル部３２および行動決定機構部３３に通知する。

【００３７】感情／本能モデル部３２は、ロボットの感情と本能の状態を表現する感情モデルと本能モデルをそれぞれ管理している。行動決定機構部３３は、センサ入力処理部３１からの状態認識情報や、感情／本能モデル部３２からの感情／本能状態情報、時間経過等に基づいて、次の行動を決定し、決定された行動の内容を行動指令情報として、姿勢遷移機構部３４に送出する。

【００３８】姿勢遷移機構部３４は、行動決定機構部３３から供給される行動指令情報に基づいて、ロボットの姿勢を、現在の姿勢から次の姿勢に遷移させるための姿勢遷移情報を生成し、これを制御機構部３５に出力する。制御機構部３５は、姿勢遷移機構部３４からの姿勢遷移情報に従って、アクチュエータ３ＡＡ$_1$乃至５Ａ$_1$および５Ａ$_2$を駆動するための制御信号を生成し、これをアクチュエータ３ＡＡ$_1$乃至５Ａ$_1$および５Ａ$_2$に送出する。これにより、アクチュエータ３ＡＡ$_1$乃至５Ａ$_1$および５Ａ$_2$は、制御信号に従って、駆動し、ロボットは、自律的に行動を起こす。

【００３９】ロボット１は、ユーザの音声とジェスチャ

を認識し、行動を決定する。図３に示した機能構成例から、ユーザの音声とジェスチャを認識し、行動を決定するための部分を取り出したものを、図４に示す。即ち、ユーザの音声を認識するために、マイク１５と音声認識部３１Ａ、ユーザのジェスチャを認識するために、ＣＣＤ１６と画像認識部３１Ｂが備えられ、音声認識部３１Ａと画像認識部３１Ｂから得られる認識結果により、行動決定機構部３３は、ロボット１の行動を決定する。

【００４０】図５は、音声認識部３１Ａの詳細な構成を示す図である。ユーザの発話は、マイク１５に入力され、マイク１５では、その発話が、電気信号としての音声信号に変換される。この音声信号は、音声認識部３１ＡのＡＤ(Analog Digital)変換部５１に供給される。ＡＤ変換部５１では、マイク１５からのアナログ信号である音声信号がサンプリング、量子化され、デジタル信号である音声データに変換される。この音声データは、特徴抽出部５２に供給される。

【００４１】特徴抽出部５２は、ＡＤ変換部５１からの音声データについて、適当なフレームごとに、例えば、スペクトルや、線形予測係数、ケプストラム係数、線スペクトル対等の特徴パラメータを抽出し、特徴量バッファ５３およびマッチング部５４に供給する。特徴量バッファ５３では、特徴抽出部５２からの特徴パラメータが一時記憶される。

【００４２】マッチング部５４は、特徴抽出部５２からの特徴パラメータ、または特徴量バッファ５３に記憶された特徴パラメータに基づき、音響モデルデータベース５５、辞書データベース５６、および文法データベース５７を必要に応じて参照しながら、マイク１５に入力された音声（入力音声）を認識する。

【００４３】即ち、音響モデルデータベース５５は、音声認識する音声の言語における個々の音素や音節などの音響的な特徴を表す音響モデルを記憶している。ここで、音響モデルとしては、例えば、HMM(Hidden Markov Model)などを用いることができる。辞書データベース５６は、認識対象の各単語について、その発音に関する情報が記述された単語辞書を記憶している。文法データベース５７は、辞書データベース５６の単語辞書に登録されている各単語が、どのように連鎖する（つながる）かを記述した文法規則を記憶している。ここで、文法規則としては、例えば、文脈自由文法（CFG）や、統計的な単語連鎖確率（N-gram）などに基づく規則を用いることができる。

【００４４】マッチング部５４は、辞書データベース５６の単語辞書を参照することにより、音響モデルデータベース５５に記憶されている音響モデルを接続することで、単語の音響モデル（単語モデル）を構成する。さらに、マッチング部５４は、幾つかの単語モデルを、文法データベース５７に記憶された文法規則を参照することにより接続し、そのようにして接続された単語モデルを用いて、特徴パラメータに基づき、例えば、HMM法等によって、マイク１５に入力された音声を認識する。そして、マッチング部５４による音声認識結果は、例えば、テキスト等で出力される。

【００４５】なお、マッチング部５４は、入力された音声を対象として、再度、処理を行うことが必要な場合は、特徴量バッファ５３に記憶された特徴パラメータを用いて処理を行うようになっており、これにより、ユーザに再度の発話を要求せずに済むようになっている。

【００４６】図６は、画像認識部３１Ｂの内部構成を示す図である。CCD１６により撮像された画像は、画像認識部３１ＢのＡＤ変換部６１に入力され、デジタルの画像データに変換され、特徴抽出部６２に出力される。特徴抽出部６２は、入力された画像データから対象物のエッジ検出や画像の濃度変化などの特徴抽出を行い、特徴パラメータまたは特徴ベクトルなどの特徴量を求める。

【００４７】特徴抽出部６２により抽出された特徴量は、顔検出部６３に出力される。顔検出部６３は、入力された特徴量からユーザの顔を検出し、その検出結果を距離測定部６４に出力する。距離測定部６４は、顔検出部６３から出力された検出結果を用いて、ユーザとの距離を測定すると共に、顔の向きを測定する。測定された測定結果は、行動決定機構部３３に出力される。

【００４８】なお、ユーザとの距離は、例えば、顔の大きさの変化から測定することが可能である。例えば、「Nerual Network-Based Frace Detection Henry A.Rowley,Shumeet Baluja, and Takeo Kanade IEEE Pattern Analysis and Machine Intellegence」に開示されている方法を用いることにより行うことが可能である。

【００４９】また、本実施の形態においては、１系統の画像入力を用いて顔の大きさの測定を行うとして説明するが、２系統の画像入力（ステレオ画像）間のマッチングを行うことによりユーザとの距離を測定しても良い。ステレオ画像からの３次元情報の抽出に関しては、例えば、「第３．３．１節ポイントパターンマッチング画像解析ハンドブック　高木幹雄、下田陽久　監修　東京大学出版会」に開示されている方法を用いることにより行うことが可能である。

【００５０】特徴抽出部６２により抽出された特徴量は、顔検出部６３に出力される一方で、マッチング部６５にも出力される。マッチング部６５は、入力された特徴量と、標準パターンデータベース６６に記憶されているパターン情報とを比較することにより得られる認識結果を行動決定機構部３３に出力する。標準パターンデータベース６６に記憶されているデータは、ジェスチャの画像データや動作パターンの特徴を示すデータである。なお、ジェスチャ認識に関しては、例えば、「ジェスチャによる感性表現の認識　日本ロボット学会誌Vol.１７　ＮＯ．７　９３３頁乃至９３６頁、１９９９」に開示されている方法を用いることが可能である。

【００５１】このように音声認識部３１Ａから出力された認識結果、および、画像認識部３１Ｂから出力された認識結果（測定結果）は、行動決定機構部３３に入力される。図７は、行動決定機構部３３の内部構成を示す図である。音声認識部３１Ａから出力された音声の認識結果は、行動決定機構部３３のテキスト解析部７１に入力される。テキスト解析部７１は、辞書データベース７２と解析用文法データベース７３に記憶されているデータを基に、入力された音声認識結果を、形態素解析、構文解析などの解析を行うことにより、単語の情報や構文の情報などの言語情報を抽出する。また、辞書に記述された内容を基に、入力の音声発話の意味、意図なども抽出する。

【００５２】すなわち、辞書データベース７２には、単語の表記や解析用文法を適用するために必要な品詞情報などの情報、単語の個別の意味情報などが記憶されており、解析用文法データベース７３には、辞書データベース７２に記憶されている各単語の情報を基に、単語連鎖に関する制約を記述したデータが記憶されている。これらのデータを用いてテキスト解析部７１は、入力された音声認識結果のテキストデータを解析する。

【００５３】解析用文法データベース７３に記憶されているデータは、正規文法、文脈自由文法、統計的な単語連鎖確立、意味的な解析までを含める場合はHPSGなどの意味論を含んだ言語理論などを用いる、テキスト解析に必要なデータである。

【００５４】テキスト解析部７１から出力された解析結果は、キーワード抽出部７４に出力される。キーワード抽出部７４は、入力された解析結果から、キーワードデータベース７５に記憶されているデータを参照して、ユーザが発話した意図を抽出し、その抽出結果を、動作表参照部７６に出力する。なお、キーワードデータベース７５には、キーワードスポッティングの際に用いられるキーワードとして、感嘆表現や命令といったユーザの意図を示す言葉のデータが保持されている。具体的には、後段の動作表参照部７６にて音声情報の索引となる表現と、それに対応した単語がキーワードのデータとして保持されている。

【００５５】動作表参照部７６は、キーワード抽出部７４から出力された抽出結果と、画像認識部３１Ｂから出力された認識結果とにより決定される動作を、動作表記憶部７７と動作分類表記憶部７８に、それぞれ記憶されている表を参照することにより決定する。ここで、動作表記憶部７７に記憶されている表について説明する。図８は、動作表記憶部７７に記憶されている動作表の一例を示す図である。

【００５６】画像の認識結果として”手招き”、”指を指す”、”握手”、”手を振る”、画像の認識結果が”ない”場合とに分類される。これらの分類によりユーザとの距離の測定結果が付帯状況として必要になってくる場合と、そうでない場合とに分けられる。さらに、音声の認識結果により、動作が決定される。

【００５７】例えば、ユーザが画像を認識した結果として”手招き”だった場合、まず、ユーザがどこにいるのか、どのくらい離れているのかなどの情報、すなわち、測定結果が必要となる。そして、ユーザが手招きしていても、その時の発話が、”こっちに来い”であれば、”ユーザに近づく”という動作が決定されるが、”向こう行け”などであれば、”ユーザから離れる”に動作が決定される。なお、詳細は後述するが、”こっちに来い”とユーザが発話した場合でも、必ず、ユーザに近づくという動作が決定されるわけではない。

【００５８】このように、動作表は、ユーザのジェスチャ（画像の認識結果）と、ユーザの発話（音声の認識結果）、さらに、状況によりユーザとの距離（測定結果）という３つの情報により、１つの動作が決定されるように記述された表である。

【００５９】図９は、動作分類表記憶部７８に記憶されている動作分類表の一例を示す図である。動作分類表は、動作表における動作を分類したものである。動作表の動作は、図９に示した動作分類表のように、４種類に分類することが可能である。すなわち、”ロボット位置相対動作”、”ユーザ位置相対動作”、”絶対位置動作”、および”その他”である。

【００６０】”ロボット位置相対動作”は、ロボットの現在位置で動作方向や距離が決定される動作であり、例えば、ユーザが”右へ行け”と発話した場合、ロボット１とユーザが対面しているときは、ユーザの右側というのは、ロボット１の左側になるので、結果として、ロボット１は、左へ移動するとする動作を行う。

【００６１】”ユーザ位置相対動作”は、ユーザの現在位置で動作方向や距離が決定される動作であり、例えば、ユーザが”こっちに来い”と発話した場合、ロボット１は、ユーザの手前８０cmのところまで行くのには、どのくらい移動したらいいかなど判断し、その判断に従って移動するという動作を行う。

【００６２】”絶対位置動作”は、ロボット１やユーザの現在位置を必要としない動作であり、例えば、ユーザの発話が”東へむかえ”であった場合、ロボット１は、東というのは、自己の位置およびユーザの位置に関係なしに決定される一意の方向なので、その方向に移動するという動作を行う。

【００６３】”その他”は、方向や距離の情報を必要としない動作であり、例えば、ロボット１が声を出すなどである。

【００６４】次に、ロボット１において行われる、ロボット１の動作の決定の仕方について説明する。上述したように、ロボット１の動作は、ユーザの音声と動作により決定される。そこで、まず、ユーザの音声を認識する動作について、図１０のフローチャートを参照して説明

する。マイク１５により取り込まれたユーザの音声は、ステップＳ１において、音声認識部３１Ａにより音声認識の処理が施される。

【００６５】音声認識部３１Ａから出力された認識結果は、ステップＳ２において、行動決定機構部３３のテキスト解析部７１に入力され、テキスト解析が行われる。そして、その解析の結果を用いて、ステップＳ３において、キーワード抽出部７４によりキーワードマッチングが行われる。その結果、キーワードが抽出されたか否かが、ステップＳ４において判断される。ステップＳ４において、キーワードが抽出されたと判断された場合、ステップＳ５に進む。

【００６６】ステップＳ５において、抽出されたキーワードを言語情報とする。一方、ステップＳ４において、キーワードは抽出されないと判断された場合、ステップＳ６に進み、キーワードは無いという情報が、言語情報とされる。ステップＳ５または、ステップＳ６の処理が終了されると、ステップＳ７において、言語情報が、動作表参照部７６に出力される。このような処理は、ロボット１が動作している間、繰り返し行われる。

【００６７】このような音声認識処理が行われる一方で、ユーザの画像に関する処理も行われている。ロボット１において行われる画像処理について、図１１のフローチャートを参照して説明する。ステップＳ１１において、CCD１６により撮像された画像は、画像認識部３１Ｂの特徴抽出部６２により特徴量が抽出される。その認識結果が用いられ、ステップＳ１２において、登録されているジェスチャがあるか否かが判断される。即ち、マッチング部６５は特徴抽出部６２により出力された特徴量を用いて、標準パターンデータベース６６に記憶されているジェスチャのパターン情報の中に、一致するものがあるか否かを判断する。このような判断により、ジェスチャがあると判断された場合、ステップＳ１３に進む。

【００６８】ステップＳ１３において、ジェスチャであると判断されたジェスチャが、付帯情報を持つものであるか否かが判断される。付帯情報を持つジェスチャとしては、例えば、ユーザが指で所定の方向をさしているような場合であり、そのような場合、その指がさしている方向に位置する対象物の情報が、付帯情報となる。ステップＳ１３において、付帯情報を持つジェスチャであると判断された場合、その付帯情報の検出が、ステップＳ１４において行われる。ステップＳ１４において、付帯情報の検出が終了されると、ステップＳ１５に進む。

【００６９】一方、ステップＳ１２において、登録されているジェスチャはないと判断された場合、又は、ステップＳ１３において、付帯情報がないと判断された場合も、ステップＳ１５の処理に進む。ステップＳ１５において、動作情報が、動作表参照部７６に出力される。

【００７０】ステップＳ１２から、ステップＳ１５の処理に進んだ場合、動作情報としては、ジェスチャはないという情報、換言すれば、画像の認識結果として動作を決定する情報はないという情報である。ステップＳ１３からステップＳ１５の処理に進んだ場合、動作情報としては、ジェスチャに関する情報のみである。ステップＳ１４からステップＳ１５の処理に進んだ場合、動作情報としては、ジェスチャに関する情報と、付帯情報である。

【００７１】このような画像認識処理は、ロボット１が動作している間、繰り返し行われる。なお、ステップＳ１３の付帯情報としては、顔検出部６３および距離測定部６４による処理の結果出力される測定結果も、必要に応じ含まれる。

【００７２】このようにして、音声認識結果としての言語情報と、画像認識結果としての動作情報を用いて、行動決定機構部３３の動作表参照部７６は、ロボット１の行動を決定する。図１２のフローチャートを参照して、動作表参照部７６の動作について説明する。ステップＳ２１において、キーワード抽出部７４から、言語情報が、画像認識部３１Ｂから動作情報が、それぞれ入力される。ステップＳ２２において、入力された言語情報と動作情報を基に、動作表記憶部７７に記憶されている動作表、および動作分類表記憶部７７に記憶されている動作分類表を参照して、一意に動作を決定する。

【００７３】ここで、決定される動作について説明する。動作は、図８に示した動作表に基づいて決定されるわけだが、例えば、画像の認識結果（動作情報）が”手招き”で、音声の認識結果（言語情報）が”こっち来い”の場合、動作としては、ユーザに近づく、ユーザから離れる、ユーザを無視するの３通りの動作が設定されている。通常なら、”手招き”されて”こっち来い”と言われれば、ユーザに近づくという動作が選択されるべきだが、常に、同じ動作を行うのでは、飽きられてしまう可能性がある。

【００７４】そこで、ユーザが同じジェスチャをし、同じ発話をした場合でも、その時々により、異なる動作を行わせるようにする。そこで、設定されている３通りの動作のうち、どの動作に決定するかは、順に決定する、ランダムに決定する、確率値により決定する、キーワードにより決定する、その時の感情により決定するなど考えられる。

【００７５】確率値により決定する場合、”近づく”が５０％、”離れる”が３０％、”無視する”が２０％というように、予め、どの動作に決定されるかの割合を決定しておく。

【００７６】キーワードにより決定する場合、現在の動作、発話と、その前の動作、発話との組み合わせにより行うことが可能である。例えば、ユーザが前の動作として手を叩き、現在の動作として、手招きしてこっちに来いといった場合、必ずユーザに近づくという動作が選択

されるように設定しておき、前の動作として殴り、現在の動作として、手招きしてこっちに来いといった場合、ユーザから離れるという動作が選択されるように設定しておく。

【００７７】このように、前の動作、発話と、現在の動作、発話の組み合わせにより動作を決定するようにしても良い。

【００７８】その時の感情により決定する場合、感情／本能モデル部３２の情報を参照し、その時の感情により、例えば、恐怖心を感じている時に、手招きしてこっちに来いといわれると、ユーザに近づき、怒りを感じている時に、手招きしてこっちに来いといわれると、ユーザを無視するといったようにすることも可能である。

【００７９】このようにして、動作表参照部７６は、言語情報と動作情報を基に、動作表を参照して動作を決定する。そして、決定された動作は、ステップＳ２３（図１２）において、姿勢遷移機構部３４に出力され、それ以降の部分で所定の処理が行われることにより、ロボット１は、決定された動作を実行する。

【００８０】上述した実施の形態においては、ユーザの指がさす方向から、ユーザの指示する方向を検出し、その方向にある対象物を付帯情報として検出するようにしたが、ユーザの顔の方向、目の向いている方向、顎のさす方向などから方向を検出し、付帯情報を検出するようにしても良い。

【００８１】また、上述した実施の形態の他に、ＯＫサイン、バツマーク、丸マーク、セーフ、耳をふさぐ（聞こえない）、威嚇、ピースサイン、まぁまぁ（手のひらを水平に揺らす）、お金を示すサイン、お願い、祈り、お手、万歳など、一般に用いられているジェスチャの情報を、標準パターンデータベース６６に記憶しておくことにより、用いることが可能となる。

【００８２】ユーザの発話したことを認識するとき、その発話自体が曖昧で（はっきりと発話されておらず）、音声認識として誤認識してしまうことがある。例えば、ユーザが”リンゴを取って”と発話したが、その発話がはっきりと発話されていないため、または、音声認識部３１Ａの誤認識の結果、”インコを取って”などと認識される場合がある。そのような場合に、画像データを用いることにより、リンゴなのかインコなのかを識別する方法について、図１３のフローチャートを参照して説明する。

【００８３】ステップＳ３１において、ユーザが発話すると、その音声はマイク１５によりロボット１に取り込まれ、音声認識部３１Ａに入力される。音声認識部３１Ａは、ステップＳ３２において、入力された音声信号を認識する。そして、その結果として、ユーザが発話したであろうと判断される複数の候補が挙げられる。それらの候補のうち、最も確からしい第１位の候補と、第２位の候補とに対し、ステップＳ３３の処理が行われる。

【００８４】ステップＳ３３において、第１位の候補と第２位の候補とのスコア値の差が、所定の閾値以内であるか否かが判断される。その結果、所定の閾値以内ではないと判断された場合、換言すれば、第１位の候補のスコア値と、第２位の候補のスコア値との差が離れているため、第１位の候補を認識結果として問題ないと判断されると、ステップＳ３７に進み、その認識結果が、音声認識結果として確定され、用いられる。

【００８５】一方、ステップＳ３３において、第１位の候補のスコア値と第２位の候補のスコア値との差が、閾値以内であると判断されると、換言すると、第１の候補が誤認識である可能性があると判断されると、ステップＳ３４に進み、スコアの高い複数の候補が、認識結果として処理対象とされる。ステップＳ３６において、画像認識が行われる。音声認識の処理対象となっているユーザの発話がされた時点で撮像された画像、または、発話がされた時点の前後の時点で撮像された画像が、ステップＳ３５における画像認識の処理対象画像である。

【００８６】ステップＳ３５における画像認識の結果を用いて、ステップＳ３６において音声認識の結果の補完が行われる。

【００８７】例えば、上述したように、ユーザが”リンゴを取って”と発話した場合、その認識結果として、第１位の候補が”リンゴを取って”であり、第２位の候補が”インコを取って”だったとする。さらに、これらの第１位の候補と第２位の候補が、所定の閾値以内であった場合、どちらの候補が正しいのか判断しかねる。そこで、その時、撮像された画像を認識し、例えば、その画像内にリンゴが撮像されていると判断される場合は、第１位の候補である”リンゴを取って”が正しい認識結果であると判断し、画像内にインコが撮像されていると判断される場合は、第２位の候補である”インコを取って”が正しい認識結果であると判断される。

【００８８】このようにして音声認識の結果が補完されると、その補完された音声認識結果が、ステップＳ３７において、音声認識結果として確定される。このように、認識結果に曖昧さが含まれるとき、画像情報を用いることにより、より確実に音声認識を行うことが可能となる。

【００８９】なお、上述した説明においては、第１位の候補と第２位の候補とのスコア値の差だけを比較したが、第１位の候補から第１０位の候補の差を取るなどの方法を用いても良い。

【００９０】ところで、ユーザＡとユーザＢが会話をしているとき、ユーザＡが”これ見て”と発話したとする。この発話に対してユーザＢが、”それはなんですか？”と発話する。このような会話は日常において、良く交わされる会話である。すなわち、同じ対象物に対しても、ユーザＡにとっては”これ”であり、ユーザＢに取っては”それ”であるというように、指示代名詞という

のは、その時の状況により異なってくるものである。

【００９１】このような事は、ユーザとロボット１が会話を行っている場合においてもいえることであり、そのために、ロボット１は、ユーザが何を指し示しているのかを、はっきりと認識する必要がある。指示代名詞が示す対象物の認識を行う時のロボット１の処理について、図１４のフローチャートを参照して説明する。ステップＳ４１において、ユーザが発話し、その発話に関してステップＳ４２において、音声認識が行われる。

【００９２】ステップＳ４３において、音声認識の結果を用いて、ユーザの発話の中に、指示代名詞があるか否かが判断される。指示代名詞がないと判断されると、その音声認識の結果が、ステップＳ４６において、音声認識結果として確定される。

【００９３】一方、ステップＳ４３において、ユーザの発話した中に指示代名詞があると判断された場合、ステップＳ４４に進み、画像認識が行われる。画像認識の対象となる画像は、ユーザが発話した時点で撮像された画像、または、ユーザが指などでさす方向を判断し、その方向にある画像を撮像した時点での画像である。

【００９４】ステップＳ４４において、撮像された画像の画像認識が行われると、ステップＳ４５において、その認識結果（画像情報）を用いて、指示代名詞の補完が行われる。ここで、具体的な例を挙げて説明する。ユーザが”それ取って”とロボット１に対して発話したとする。その際、ユーザは、”それ”に対応する対象物を指で指し示すなどのジェスチャをとる。

【００９５】ロボット１は、その発話を受け、ステップＳ４２において、音声認識を行い、その結果、指示代名詞である”それ”を含んでいると判断する。また、ユーザが発話した時点で撮像された画像から、ユーザが指で所定の方向を指し示すというジェスチャをとっているということが判断される。

【００９６】ステップＳ４４において、ロボット１は、ユーザが”それ”と指し示した方向を判断し、その方向の画像を撮像し、撮像された画像に対して画像認識を行う。その画像認識の結果、例えば、新聞が対象物として認識されると、”それ”という指示代名詞が示す対象物は”新聞”であると補完される。このようにして、ステップＳ４５において、画像情報から指示代名詞が補完されると、ステップＳ４６に進み、その補完された音声認識結果が、音声認識結果として確定される。

【００９７】このように、画像情報を用いることにより、指示代名詞が示す対象物を、確実に認識することが可能となる。

【００９８】ロボット１が画像を撮像した場合、その撮像された画像内には、複数の物体を含んでいる事が多い。そのような複数の物体の内、ユーザが会話の対象物して指し示している物体は何であるかを認識する処理について、図１５のフローチャートを参照して説明する。ステップＳ５１において、ＣＣＤ１６により撮像されたユーザのジェスチャが画像としてロボット１に入力される。

【００９９】ジェスチャが入力されることにより、例えば、そのジェスチャが所定の方向を指し示すといったものであった場合、付帯情報を検出するために、ユーザが指し示す方向の画像を認識する必要がある。そこで、ユーザが指し示す方向の画像が撮像され、その画像に関して、画像認識部３１Ｂによる画像認識処理が、ステップＳ５２において行われる。その認識結果が用いられ、ステップＳ５３において、複数の対象物が、画像内に存在するか否かが判断される。ステップＳ５３において、対象物は複数存在しないと判断されると、換言すれば、対象物として１つしか存在しないと判断されると、ステップＳ５６に進み、その対象物の画像認識結果が、画像認識結果として出力される。

【０１００】一方、ステップＳ５３において、対象物が複数あると判断された場合、ステップＳ５４に進み、音声認識が行われる。音声認識の対象となる音声は、ユーザがジェスチャを行った時点で取り込まれた音声である。ステップＳ５４における音声認識の結果（音声情報）が用いられ、ステップＳ５５において、画像認識結果の補完が行われる。ここで、具体例を挙げて説明する。

【０１０１】ユーザが、所定の方向を指し示すというジェスチャをしながら、”ボールを取って”と発話したとする。まずロボット１は、ユーザのジェスチャを認識し、そのジェスチャが、所定の方向を指し示すジェスチャであると認識する。そして、その指し示す方向の画像を撮像し、画像内の対象物の認識を行う。その結果、複数の対象物が存在すると判断されると、ジェスチャと同時にユーザが発話した音声の認識を行う。

【０１０２】その音声認識の結果、”ボールを取って”と認識されると、そのうちの”ボール”が、画像内の複数の対象物のなかでも、ユーザが所望としている対象物であると判断される。すなわち、音声情報から画像認識結果が補完される。このようにして音声情報から画像認識結果が補完されると、ステップＳ５６に進み、補完された画像認識結果が、画像認識結果として出力される。

【０１０３】このようにして、音声情報を用いて、画像情報の曖昧な部分を補完することにより、より精度の高い画像情報の取得を行う事が可能となる。

【０１０４】ところで、音声情報のみにより行動を起こすロボットは、例えば、ユーザの音声がする方向に進むといった行動し、画像情報のみにより行動を起こすロボットは、例えば、撮像している画像内にユーザが収まる方向に進といった行動を行う。しかしながら、本発明を適用したロボット１は、上述したように、音声情報と画像情報とを組み合わせて、ユーザが求めている動作を判断し、実際に行動に移す。そこで、ロボット１の行動を

分類すると、既に説明したように、図９に示したような動作分類表のように分類することができる。

【０１０５】すなわち、音声を認識し、画像情報からユーザやロボット１自身の位置を把握することにより、動作が決定される。具体的には、ユーザが”こっちに来い”と発話したとき、まず、その発話が音声認識され、次に、ユーザの位置を画像情報から認識する。そして、ユーザの方向に進むという動作が決定された場合、どれだけの距離を、どの方向に進むかという目的位置が決定される。

【０１０６】例えば、図１６に示すように、目的位置としては、ユーザの手前８０ｃｍの所と設定される。これは、画像認識部３１Ｂの特徴抽出部６２（図６）が抽出した特徴量を用いて、顔検出部６３がユーザの顔を認識し、距離測定部６４により、認識されたユーザの顔の大きさを基に、ユーザと自己の間の距離が測定される。そして、その測定された距離を用いて、ユーザの手前８０ｃｍまで移動するには、どれだけ進めば良いかが決定される。

【０１０７】このように、ユーザの位置の測定を行い、動作に応じて、ユーザの位置を利用することにより、ユーザのジェスチャに対する動作を、より正確なものにすることが可能となる。

【０１０８】上述した実施の形態においては、ユーザの音声として実際に発話した言葉を認識するようにしたが、音声として、ユーザが手拍子で出した音（リズム）や、ユーザの足音（音響）を用いて、ロボット１の動作を決定するようにすることもできる。

【０１０９】ユーザのリズムや音響（以下、適宜、これらを含めたものを音声と表現する）を用いる場合、音声認識部３１Ａの構成は、図１７に示したような構成となる。すなわち、マイク１５により取り込まれた音声は、ＡＤ変換部５１に入力され、デジタルデータに変換され、さらに、リズム／音響認識部８１に入力される。リズム／音響認識部８１は、リズムや音響に関する情報を取得する。

【０１１０】リズム／音響認識部８１により取得された認識結果は、行動決定機構部３３に出力される。なお、図１７には、ユーザの発話を認識する部分、すなわち、図５に示した部分に関しては、省略して記述してある。従って、ＡＤ変換部５１から出力されたデジタルの音声信号は、特徴抽出部５２（図５）に出力されるとともに、リズム／音響認識部８１（図１７）にも出力される。

【０１１１】さらに、リズム／音響認識部８１から出力された認識結果は、行動決定機構部３３に出力されるが、行動決定記後部３３のテキスト解析部７１（図７）に入力されるのではなく、直接、動作表参照部７６に入力される。

【０１１２】ここで、リズム／音響認識部８１が行うリズムの認識方法について説明する。リズム／音響認識部８１打楽器音（ユーザの手拍子による音を含む）のビート（拍）検出、あるいは、コード（和音）変化によるビート検出などの方法を用いてリズムの検出を行う。その結果、いつビートが検出されたか、何拍子か、１分あたり何拍かという検出結果が出力される。

【０１１３】リズム検出の方法に関しては、「打楽器音を対象にした音源分離システム、後藤真考、村岡洋一著、電子情報通信学会論文誌、J77-DII、NO.5、901乃至911頁、1994」や、「音響信号を対象としたリアルタイムビートトラッキングシステム、後藤真考、村岡洋一著、電子情報通信学会論文誌、J81-DII、NO.2、227乃至237頁、1998」などに開示されている方法を用いることが可能である。

【０１１４】リズム／音響認識部８１から出力されたリズムに関する認識結果を用いて、ここでは、行動決定機構部３３（動作表参照部７６）が決定する動作として踊りを踊る場合を例に挙げて説明する。動作表記憶部７７には、図１８に示したような動作表が記憶されている。例えば、リズムに関する認識結果が、１分間に０乃至６０拍であり、２拍子の場合、踊りＡが選択され、１分間に０乃至６０拍であり、２拍子でも、３拍子でも、４拍子でもない場合、踊りＡが選択されというように、１分間の拍と、何拍かという情報により、一意に踊りのタイプが選択される。

【０１１５】このようにして動作表参照部７６が動作表記憶部７７に記憶されている動作表を参照することにより決定された動作が行われるように、行動決定機構部３３以降の部分で所定の処理が行われることにより、ロボット１は制御される。

【０１１６】上述した説明では、音声によりリズムに関する情報を取得するようにしたが、ジェスチャにより取得されるようにしても良い。ジェスチャによりリズムに関する情報を取得するようにした場合、画像認識部３１Ｂの構成は、図６に示したような構成で良い。ジェスチャによりリズムに関する情報を取得する方法としては、「ジェスチャによる感性表現の認識、井口征士著、日本ロボット学会誌１７巻７号」に開示されている方法を用いることが可能である。

【０１１７】もちろん、音声およびジェスチャの両方から、リズムに関する情報を取得するようにしても良い。

【０１１８】次に、ロボット１の動作が音響により決定される場合について説明する。リズム／音響認識部８１により認識される音響としては、例えば、足音か、悲鳴かなど、どのような種類の音であるのか、また、好きな人が発した音、嫌いな人が発した音、車が発した音など、誰が、あるいは、何が発した音なのかを示すものである。

【０１１９】リズム／音響認識部８１において認識された結果は、動作表参照部７６に出力される。動作表参照

部７６は、動作表記憶部７７に記憶されている動作表を参照して、入力された音響に関する認識結果に対応する動作を決定する。動作表記憶部７７に記憶されてる音響に関する動作表の一例を、図１９に示す。

【０１２０】図１９に示した動作表としては、例えば、音響の認識結果として、足音が検出され、その足音が、好きな人の足音であると判断されると、喜びなら近づくという動作が選択されというように、音響現象により一意に行動が決定されるような表である。好きな人、嫌いな人という情報は、ユーザとロボット１との間で交わされる会話や、ユーザの態度などからロボット１が判断し、情報として記憶するようにしてもよい。

【０１２１】また、音響だけでなく、画像情報を用いるようにしても良い。すなわち、足音が聞こえたとき、その足音から誰か来たのかを判断することも可能であるが、画像として撮像され、認識された結果から、誰だかを判断し、その人が好きな人なのか、嫌いな人なのかを判断し、動作を決定するようにしても良い。

【０１２２】上述したように、音声情報と画像情報を組み合わせることにより、さまざまな動作をロボット１に行わせる事が可能となり、さらに、その動作決定における音声および画像の認識の段階において、互いの情報を用いることにより、より制度の高い認識処理を行う事が可能となる。

【０１２３】上述した一連の処理は、ハードウェアにより実行させることもできるが、ソフトウェアにより実行させることもできる。一連の処理をソフトウェアにより実行させる場合には、そのソフトウェアを構成するプログラムが専用のハードウェアに組み込まれているコンピュータ、または、各種のプログラムをインストールすることで、各種の機能を実行することが可能な、例えば汎用のパーソナルコンピュータなどに、記録媒体からインストールされる。

【０１２４】この記録媒体は、図２０に示すように、コンピュータとは別に、ユーザにプログラムを提供するために配布される、プログラムが記録されている磁気ディスク１３１（フロッピディスクを含む）、光ディスク１３２（CD-ROM（Compact Disk-Read Only Memory），DVD（Digital Versatile Disk）を含む）、光磁気ディスク１３３（MD（Mini-Disk）を含む）、若しくは半導体メモリ１３４などよりなるパッケージメディアにより構成されるだけでなく、コンピュータに予め組み込まれた状態でユーザに提供される、プログラムが記憶されているROM１１２や記憶部１１８が含まれるハードディスクなどで構成される。

【０１２５】なお、本明細書において、媒体により提供されるプログラムを記述するステップは、記載された順序に従って、時系列的に行われる処理は勿論、必ずしも時系列的に処理されなくとも、並列的あるいは個別に実行される処理をも含むものである。

【０１２６】また、本明細書において、システムとは、複数の装置により構成される装置全体を表すものである。

【０１２７】

【発明の効果】以上の如く請求項１に記載の情報処理装置、請求項１０に記載の情報処理方法、および請求項１１に記載の記録媒体によれば、音声を認識し、画像を認識し、音声の認識結果、または、画像の認識結果のうち、少なくとも一方を用いて、ロボットの動作を決定するようにしたので、より制度の高い音声認識および画像認識を行うことが可能となる。

【図面の簡単な説明】

【図１】本発明を適用したロボットの一実施の形態の外観構成例を示す斜視図である。

【図２】図１のロボットの内部構成例を示すブロック図である。

【図３】図２のコントローラ１０の機能的構成例を示すブロック図である。

【図４】音声および画像を認識し行動を決定する部分に関する機能構成例を示す図である。

【図５】音声認識部３１Ａの内部構成を示すブロック図である。

【図６】画像認識部３１Ｂの内部構成を示すブロック図である。

【図７】行動決定機構部３３の内部構成を示すブロック図である。

【図８】動作表記憶部７７に記憶されている動作表について説明する図である。

【図９】動作分類表記憶部７８に記憶されている動作分類表について説明する図である。

【図１０】音声認識処理について説明するフローチャートである。

【図１１】画像認識処理について説明するフローチャートである。

【図１２】動作決定処理について説明するフローチャートである。

【図１３】音声情報と画像情報とを用いて、認識結果を出力する場合の処理について説明するフローチャートである。

【図１４】音声情報と画像情報とを用いて、認識結果を出力する場合の他の処理について説明するフローチャートである。

【図１５】音声情報と画像情報とを用いて、認識結果を出力する場合のさらに他の処理について説明するフローチャートである。

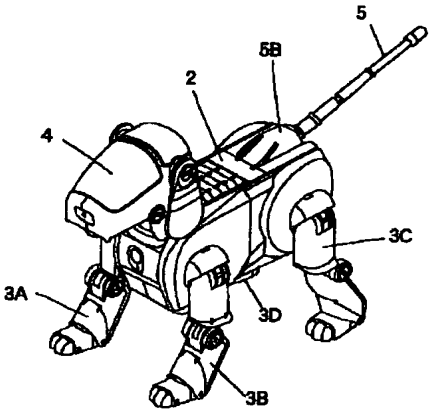
【図１６】ユーザとロボット１の位置関係について説明する図である。

【図１７】音声認識部３１Ａの他の構成を示す図である。

【図１８】動作表記憶部７７に記憶されている他の動作

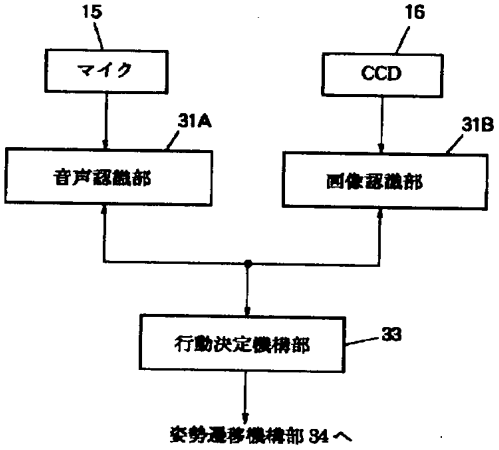
表について説明する図である。
【図１９】動作表記憶部７７に記憶されている、さらに他の動作表について説明する図である。
【図２０】媒体を説明する図である。
【符号の説明】
１０　コントローラ，　１０Ａ　CPU，　１０Ｂ　メモ＊リ，　１５　マイク，１６　CCD，　１７　タッチセンサ，　１８　スピーカ，　１９　通信部，３１　センサ入力処理部，　３１Ａ　音声認識部，　３１Ｂ　画像認識部，　３１Ｃ　圧力処理部，　３２　感情／本能モデル部，　３３　行動決定機構部，３４　姿勢遷移機構部，　３５　制御機構部，　３６　音声合成部

【図１】

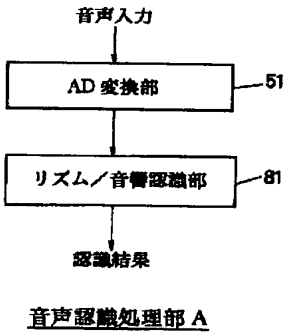


ペットロボット１の外観構成

【図４】

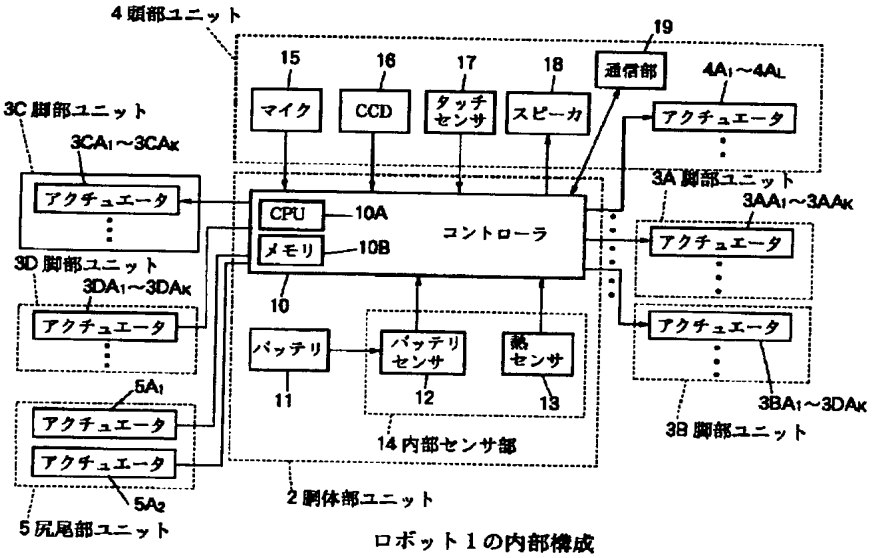


【図１７】



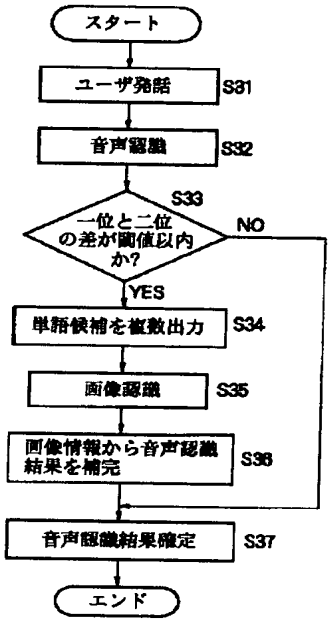
音声認識処理部Ａ

【図２】

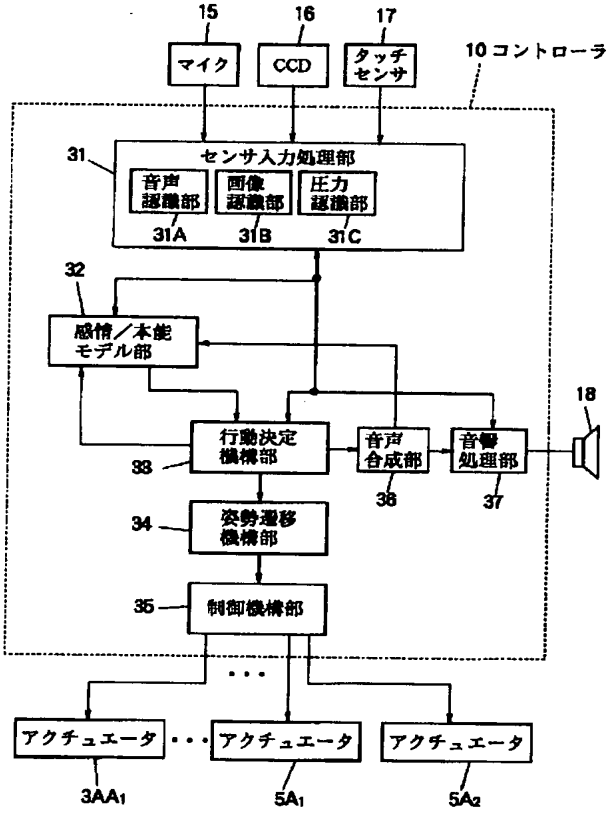


ロボット１の内部構成

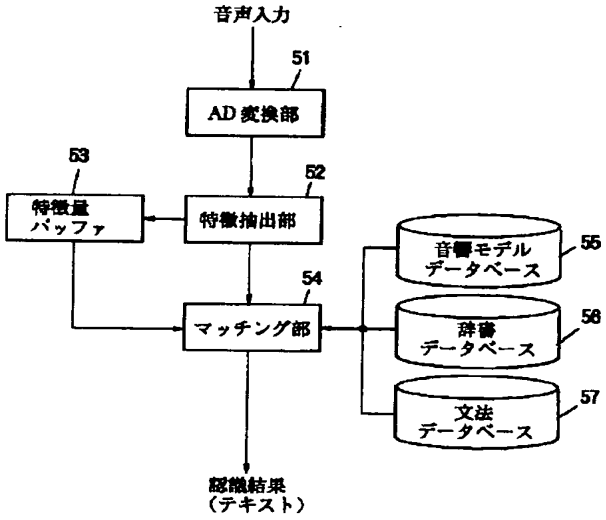
【図１３】



【図３】

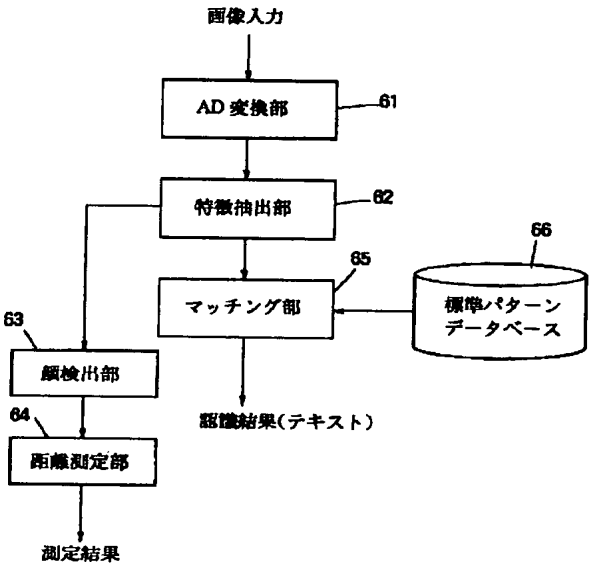


【図５】



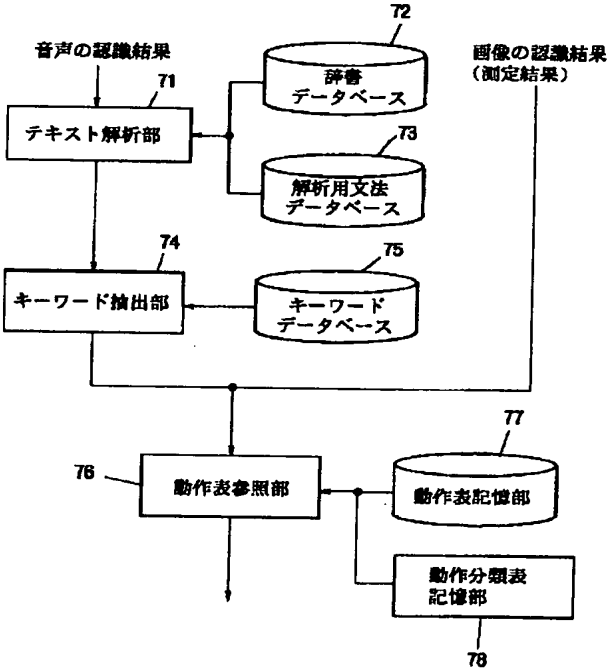
音声認識部 31A

【図６】



画像認識部 31B

【図７】



行動決定機構部 33

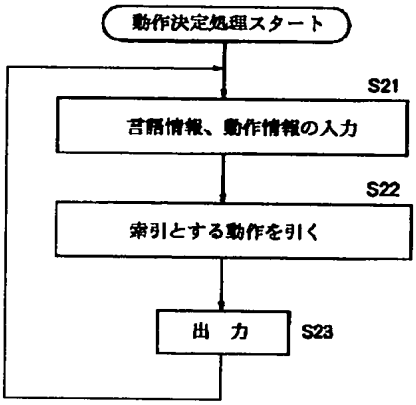
【図８】

画像の認識結果	付帯情報	音声の認識結果		動作
手招き	ユーザの位置	こっち来い等 ロボットの名前(名前を呼ぶ)		ユーザに近づく、離れる、無視
ゆびを指す	指の方向	感嘆表現(あっ、おやっ等)		指の差している方向を向く
		～を取って	指示代名詞	指の指す方向にある対象物を切り出し指示代名詞を補完
			物の名前	ゆびの指す方向で物を検索する
握手		挨拶		手を前に出す
手を振る		別れの挨拶(バイバイ等)		手を振る、ユーザーから離れる、電源を切る
		挨拶(おーい等呼びかけ)		近づく、挨拶
		なし		手を振る
なし		呼びかけ		ユーザーを探す

動作表

77 動作表記憶部

【図１２】



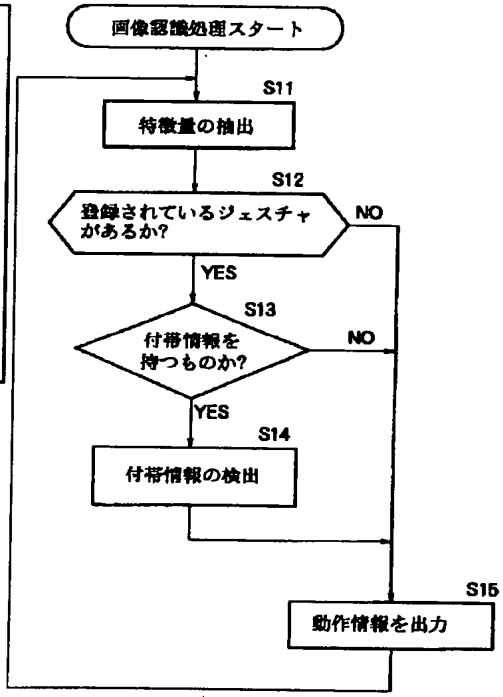
【図９】

ロボット位置相対動作
「右へ行け」、「さがれ」等のロボットの現在位置で動作方向や距離が決定できる動作。
ユーザ位置相対動作
「こっちへ来い」、「あっちへいけ」等、動作方向や距離の決定にユーザ位置を必要とする動作
絶対位置動作
「東へむかえ」等の、ロボットとユーザの現在位置を必要としない動作。
その他
ロボットが声を出す等の、方向や距離位置を必要としない動作。

動作分類表

78 動作分類表記憶部

【図１１】



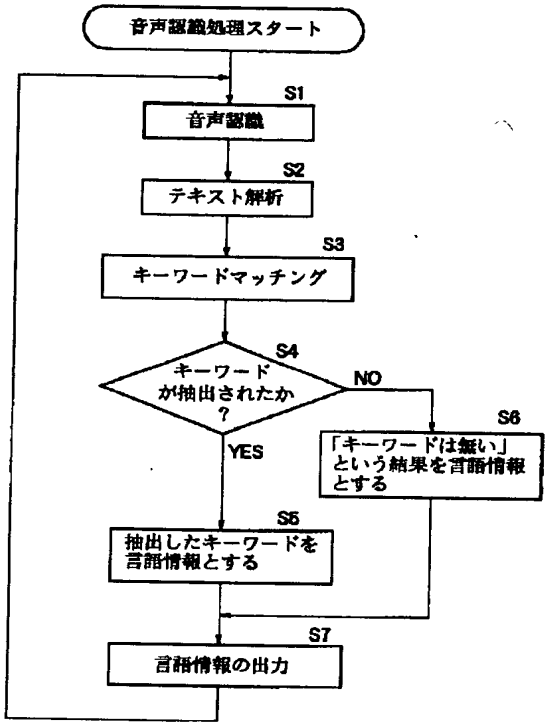
【図１８】

	2拍子	3拍子	4拍子	それ以外
0 － 60	踊りA	踊りA	踊りA	踊りA
70 －120	踊りB	踊りC	踊りD	踊りA
120 －180	踊りE	踊りF	踊りG	踊りA

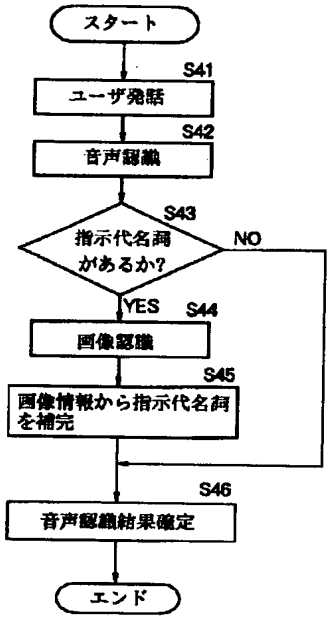
動作表

77 動作表記憶部

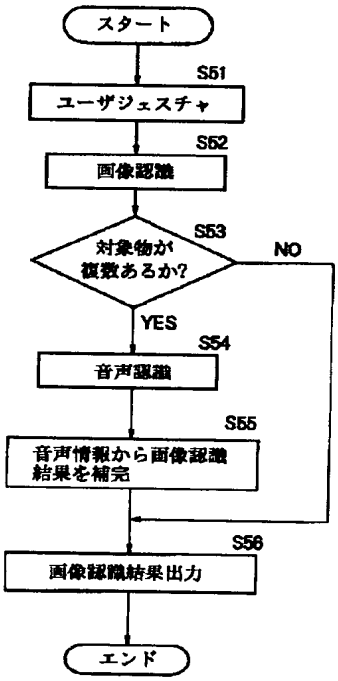
【図１０】



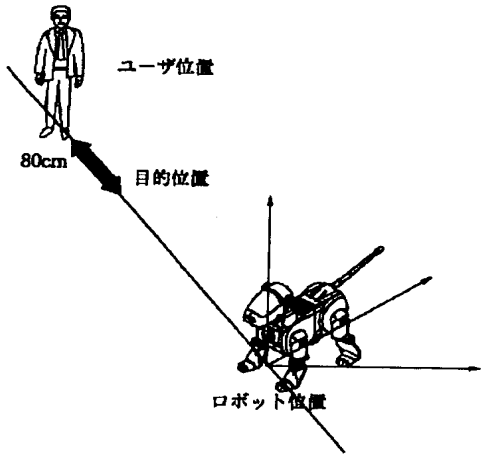
【図１４】



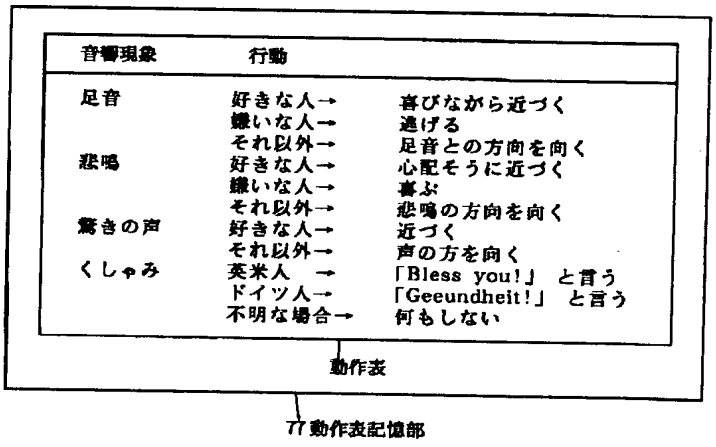
【図１５】



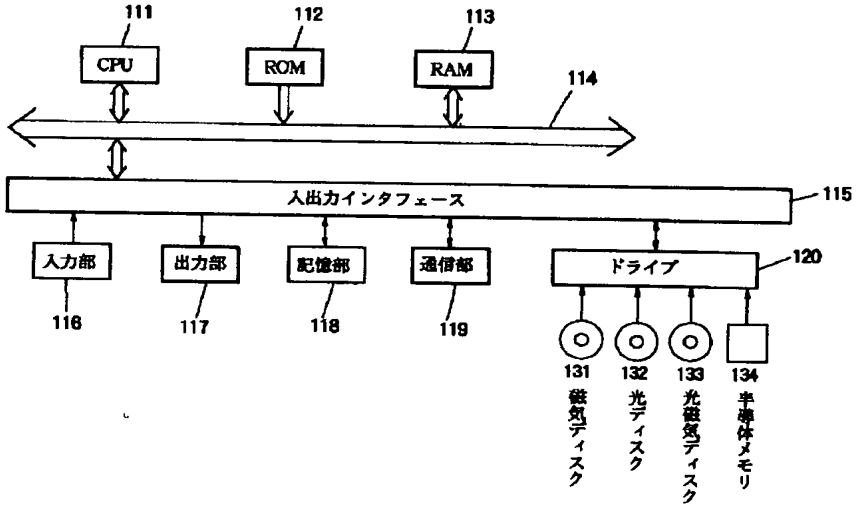
【図１６】



【図１９】



【図２０】



フロントページの続き

(72)発明者　本田　等
東京都品川区北品川６丁目７番35号　ソニー株式会社内
(72)発明者　ヘルムート　ルッケ
東京都品川区北品川６丁目７番35号　ソニー株式会社内
(72)発明者　田丸　英司
東京都品川区北品川６丁目７番35号　ソニー株式会社内
(72)発明者　藤田　八重子
東京都品川区北品川６丁目７番35号　ソニー株式会社内

Ｆターム(参考)　3F059 AA00 BA02 BB07 BC06 CA05
CA06 DA02 DA03 DA05 DA09
DB02 DB09 DC01 DC04 DC07
DD01 DD06 DD18 FA03 FA05
FB01 FC07 FC14 FC15
5B057 AA05 BA02 CA08 CA12 CA16
DA17 DB02 DB09 DC08 DC16
DC22 DC36
5D015 KK01 LL07
5L096 AA06 CA14 DA02 FA06 FA14
FA66 FA67 HA03
9A001 HH17 HH19 HH20